MEASURING SEMANTIC RELATEDNESS USING SALIENT ENCYCLOPEDIC CONCEPTS

Samer Hassan

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 2011

APPROVED:

Rada Mihalcea, Major Professor
Paul Tarau, Committee Member
Miguel Ruiz, Committee Member
Andras Csomai, Committee Member
Ian Parberry, Interim Chair of the Computer Science and Engineering Department
Costas Tsatsoulis, Dean of the College of Engineering
James D. Meernik, Acting Dean of the Toulouse Graduate School

Hassan, Samer. Measuring Semantic Relatedness Using Salient Encyclopedic Concepts. Doctor of Philosophy (Computer Science), August 2011, 100 pp., 22 tables, 10 illustrations, 92 references.

While pragmatics, through its integration of situational awareness and real world relevant knowledge, offers a high level of analysis that is suitable for real interpretation of natural dialogue, semantics, on the other end, represents a lower yet more tractable and affordable linguistic level of analysis using current technologies. Generally, the understanding of semantic meaning in literature has revolved around the famous quote ``You shall know a word by the company it keeps''. In this thesis we investigate the role of context constituents in decoding the semantic meaning of the engulfing context; specifically we probe the role of salient concepts, defined as content-bearing expressions which afford encyclopedic definitions, as a suitable source of semantic clues to an unambiguous interpretation of context. Furthermore, we integrate this world knowledge in building a new and robust unsupervised semantic model and apply it to entail semantic relatedness between textual pairs, whether they are words, sentences or paragraphs. Moreover, we explore the abstraction of semantics across languages and utilize our findings into building a novel multi-lingual semantic relatedness model exploiting information acquired from various languages. We demonstrate the effectiveness and the superiority of our mono-lingual and multi-lingual models through a comprehensive set of evaluations on specialized synthetic datasets for semantic relatedness as well as real world applications such as paraphrase detection and short answer grading. Our work represents a novel approach to integrate world-knowledge into current semantic models and a means to cross the language boundary for a better and more robust semantic relatedness representation, thus opening the door for an improved abstraction of meaning that carries the potential of ultimately imparting understanding of natural language to machines.

ACKNOWLEDGMENTS

While all of us strive to be perfect, to be smarter, to be wiser, and to be more patient, the gods of character has been more generous to some over others. Specifically, I would like to thank my exemplary advisor, Dr. Rada Mihalcea, for her unyielding support. Aside from being a superb researcher, she has been a magnificent educator, a mentor, and a friend who was able to spark my interest in this field and guided me gracefully through this road. I and many others in our research group consider her as a super human for her wonderful qualities that we all aspire to. Additionally, I would like to acknowledge my wonderful wife, Carmen Banea who is, in every way, my better half. Without her patience, kindness, feedback, and support this road would have been a very difficult one. I consider myself very fortunate to be in the company of such a perfect person and I only hope that perfection is contagious. Similarly, I would like to acknowledge my family, especially my mother for seeding my curiosity for learning and exposing me to literature from an early age, for her unyielding love and care which helped me become what I am today.

Not to forget my friends who tolerated my sense of humor and lighted my days, who shared my good and bad moments, and who were always there to help. At last, I would like to thank my PhD committee for spending their valuable time and effort to guide me through my research and dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1. Problem Definition

Semantic relatedness is the task of finding and quantifying the strength of the semantic connections that exist between textual units, be they word pairs, sentence pairs, or document pairs. It is one of the main tasks explored in the field of natural language processing, as it lies at the core of a large number of applications such as information retrieval [69], query reformulation [10, 57, 78, 90], image retrieval [23, 48], plagiarism detection [8, 9, 27, 29, 53, 83], information flow [57], sponsored search [10], short answer grading [62, 64, 72], and textual entailment [15].

For instance, one may want to determine how semantically related are "car" and "automobile", or "noon" and "string". Similarly, one may want to find the relatedness of two pieces of text such as "I love animals" versus "I own a pet." To make such judgments, humans typically rely on accumulated knowledge and experiences, and utilize their ability of conceptual thinking, abstraction, and generalization. The importance of conceptual thinking and semantic abstraction becomes clear in neuroscience research by studying the deficiency of such ability in human subjects with developmental disorders, namely autistic savants [73]. Autistic savants have extraordinary artistic, spatial, musical, and mathematical skills accompanied by an uncanny ability of acquisition and recall of meaningless raw details - memory without understanding [77]. In a word association experiment, the autistic savants are presented with a list of words that relate very strongly to the $sweet$ concept (e.g. "sugar," "candy," "chocolate," etc.). When questioned about the list upon inclusion of new words such as "sweet," the autistic savants are able to recognize that this term was not in the original list, while control subjects generally overlook this fact. The key reason is that the human mind typically conceptualizes related terms and hence erroneously recognizes "sweet" as a previously cited term, unlike the brain circuitry of autistic savants who enables them to retain raw information without any abstraction. Accordingly, it seems that for any system to achieve a

higher conceptual understanding, the system should not only be able to acquire and use a large background knowledge [47], it should also able to abstract and generalize it.

1.2. Proposed Solution

> You shall know a word by the company it keeps.
>
> *J. R. Firth*

While Firth's notion of word meaning depends on the word's context, there have been various interpretations of what constitutes a context. Previous work has limited the context to a specific part-of-speech (e.g. verbs) [67], a part-of-speech grouping (e.g. nouns, verbs, and adjectives) [17], or the entire vocabulary [31]. More recent work has adopted word senses rather than raw words to resolve context ambiguity [63], however such techniques are difficult to scale due to the absence of large-scale sense annotated corpora. In this work, I introduce a novel and robust interpretation of the context notion by attributing to it a set of well defined and unambiguous "concepts," where a concept is a term that can afford an encyclopedic definition. In this interpretation, a casual reader's representation of a given text is modeled as a mental frame that retains and pairs unambiguously defined terms. Subsequently, this allows for an easy anchoring and association with other relevant mental knowledge, as well as easier recall. To implement this model I utilize annotated encyclopedic resources like Wikipedia, where salient concepts within each article are tagged (by being marked as hyperlinks). The annotation is further expanded by utilizing word sense disambiguation heuristics, hence generating a large-scale corpus annotated for saliency. Further analysis of the generated corpus yields semantic vectors representing words' meaning in a concept-space. Such a deviation from a simple word-space to a richer and concrete concept-space model corresponds closely to what Lenat & Feigenbaum [46] refer to as the knowledge principle.

> If a program is to perform a complex task well, it must know a great deal about the world it operates in.
>
> *Lenat and Feigenbaum*

Indeed, encyclopedic knowledge represents a window to world knowledge and by incorporating such knowledge into our semantic models we arrive one step closer to simulating natural language understanding by machines.

Due to the diversity of the semantic knowledge encoded into different languages, I also seek to explore the abstraction of semantic relatedness across languages. Through human annotation experiments, I analyze whether semantic relatedness can be carried across multiple languages with minimal dilution and utilize the findings to build an original multilingual semantic relatedness model exploiting information acquired from various languages. This work represents a novel approach to integrate world knowledge into current semantic models and a means to cross the language boundary for a better and more robust semantic relatedness representation, thus opening the door for an improved abstraction of meaning that carries the potential of ultimately imparting understanding of natural language to machines.

1.3. Contributions

The contributions of this work encompass multiple facets ranging from incorporating word knowledge into semantic models to stepping across language boundaries for a richer and more robust semantic relatedness representation. They are enumerated as follows:

i. **Propose a new interpretation of the semantic context using salient concepts identified in a context's lexicon.**

   Identifying the correct semantic interpretation of context is very valuable in understanding natural language. Therefore this thesis investigates the role of a special lexicon tagged as "concepts" in the semantic modelling of context and the way this can be interpolated to generate a semantic interpretation of all lexical units may they be words, sentences, or documents.

ii. **Examine the salient concept-space representation of meaning as more effective and robust when compared to contending representations.**

   Since the literature offers various examples of semantic representation, I investigate the advantages and disadvantages of the concept-space representation when compared to traditional word-space and lexical-knowledge models. In particular, I examine the role of the rich concept-space representation employed by this model, in contrast to similar concept-space representations adopted by competing models such as Latent Semantic Analysis and Explicit

Semantic Analysis, as well as lexical-knowledge employing human expertise in constructing ubiquitous semantic abstraction.

iii. **Explore the association between the semantic relatedness of context and the choice of communicating language.**

Here, I explore the portability of semantic relatedness across languages under controlled settings and whether it is affected by the choice of the target language or the subjectivity of the annotator. For example, could the semantic relatedness between two textual units be transferred from one language to another without perturbing its intended use? This entails overcoming the ambiguity inherent in the source language, the ambiguity imposed by the target language, as well as the subjective interpretation of the translator/annotator.

iv. **Propose a new scheme of incorporating mono-lingual semantic models in a multilingual setting to improve semantic relatedness.**

Under the hypothesis that semantics represents a higher level of abstraction that is independent of the choice of the underlying language, I seek to answer the question of whether incorporating additional semantic clues from parallel multilingual resources can improve the semantic relatedness task. In particular, can mono-lingual semantic relatedness models from different languages be aggregated to induce an overall stronger and more coherent semantic relatedness of contexts?

v. **Propose a framework for evaluating semantic relatedness in multilingual settings.**

Since this research requires the utilization of multiple monolingual semantic models covering a diverse set of languages, it is paramount to establish a robust evaluation framework to accommodate these settings. This includes formally defining and standardizing the process of dataset construction, the evaluation metrics, and the assessment strategy.

1.4. Thesis Outline

The thesis is organized as follows. Chapter 2 presents a background on semantic relatedness and its applications. Chapter 3 covers the related work in this domain. Chapter 4 introduces the proposed model and its implementation details. Chapters 5 and 6 address the experimental

setup and evaluations, respectively. Finally, I conclude with Chapter 7 and frame potential future work directions.

CHAPTER 2

BACKGROUND

This chapter addresses some of the terminologies used throughout the thesis and discusses some real-life application of semantic relatedness.

2.1. Syntax, Semantics, and Pragmatics

While words represent the building blocks of natural language, *syntax* deals with the rules (grammar) that govern it; hence syntax represents order and is a prerequisite for understanding natural language. For example, the sentence "man bites dog" is very different in syntax and meaning from the sentence "dog bites man," even though both sentences share the same lexicon. While syntax implies order and structure, *semantics* deals with the mental interpretation of natural language, much like interpreting the meaning of a painting which is constructed from an arrangement of colors and lines to convey a message. Semantics relies on analyzing bonds and relations between words, phrases, and sentences to uncover explicit or implicit meanings and interpretations. Hence, it enables natural language to be a medium through which knowledge and information are exchanged. *Pragmatics* integrates contextual awareness (situation) and incorporates non-linguistic clues such as (but not limited to) author intent, mode, and utterance. These additional clues allow further disambiguation of meaning and coherent semantic interpretation. For example, continuing the painting metaphor, its meaning may be better grasped by knowing important life events about the painter, his or her character, personality and emotional state, and the historical context of the painting, all of which may not be directly observed within the painting itself.

2.2. Semantic Relatedness vs. Semantic Similarity

A difference is usually made between semantic *relatedness* and semantic *similarity* [11]. *Similarity* is a more specific concept than *relatedness*: similarity is concerned with entities related by virtue of their likeness which share the same part-of-speech, such as *error-mistake* (near

synonym) or *fish-animal* (hypernym). On the other hand, semantic relatedness is more general and covers a wider range of relationships between entities. e.g., *hot-cold* (antonym), *fast-train* (collocation), *hiking-mountain* (association), and *car-wheel* (meronym/part-of).

2.3. Semantic Relatedness Applications

Semantic relatedness lies at the core of a large number of applications such as information retrieval, query reformulation, image retrieval, plagiarism detection, information flow, sponsored search, paraphrase detection, short answer grading, and textual entailment.

From these applications, I focus on paraphrase detection and short answer grading as an evaluation test bed.

2.3.1. Paraphrase Detection

> "The Iraqi Foreign Minister warned of disastrous consequences if Turkey launched an invasion of Iraq."
>
> "Iraq has warned that a Turkish incursion would have disastrous results."

Paraphrasing can be modelled as the rearticulation of a given text while preserving the original meaning. Accordingly, paraphrase detection is the task of automatically identifying possible paraphrases given a pair of candidate texts. Since there are no linguistic constraints on the formulation of paraphrases, the paraphrases may not share any lexicon. Additionally, non-paraphrases might share a significant part of their lexicon. For example, a simple introduction of negation would be sufficient to break possible paraphrases. Accordingly, automatic paraphrase detection is not a trivial task and it requires much more than just a bag-of-words approach [1]. Automatic paraphrase detection is useful in many domains, however it is especially valuable in plagiarism detection where it can automatically identify and quantify the degree of semantic and lexical overlap between sources.

2.3.2. Short Answer Grading

Question: What are the main advantages associated with object-oriented programming?

Correct answer: Abstraction and re-usability.

Student answer: They make it easier to reuse and adapt previously written code and they separate complex programs into smaller, easier to understand classes.

Short answer grading is typically a demanding task which requires the involvement of human graders. In this task, graders have to score the "relatedness" of the student's answer to the correct answer key. Due to the relative subjectivity of this task, human graders usually demonstrate moderate correlation (0.64-0.74) [64]. Many automated models have been utilized in this task ranging from bag-of-words overlap to pattern-matching. Recent models incorporate semantic relatedness, thus allowing for faster and more robust annotations as demonstrated in Mohler & Mihalcea [64] and Wiemer-Hastings et al. [88].

2.4. Wikipedia

Wikipedia is a free multilingual collaborative encyclopedic resource which allows volunteer users to continuously introduce, annotate, and improve current knowledge. The basic building block of Wikipedia is an *article*. Each article defines and explains a concept (entity or event). Throughout the explanations, citations to other relevant concepts are typically hyperlinked to the corresponding Wikipedia article, hence prompting the reader to inquire about and explore these relevant concepts in a convenient and coherent fashion. This flexibility led to a tremendous increase of documented knowledge in the form of high quality up-to-date articles. This is evident from a comparative study performed by Giles [22] which concludes that Wikipedia scientific contents are almost as accurate as the Encyclopedia Britannica.

The size of Wikipedia grows at a very fast pace. To illustrate, the English Wikipedia grew from 30 articles in 2001, to $438,000$ articles in 2005, and more than $2,700,000$ articles in 2009. Likewise, the Chinese Wikipedia grew from 75 articles in 2002, to $45,000$ articles in 2005, to $205,000$ articles in 2009.

Wikipedia editions are available in more than 250 languages, with the number of entries varying from a few pages to more than three millions articles per language[1]. Table 2.1 shows the

[1]In the reported experiments, I use a download from October 2008 of the English Wikipedia, with approximately 6 million pages, and more than 9.5 million hyperlinks.

TABLE 2.1. Top ten largest Wikipedias

Language	Articles	Users
English	2,221,980	8,944,947
German	864,049	700,980
French	765,350	546,009
Polish	579,170	251,608
Japanese	562,295	284,031
Italian	540,725	354,347
Dutch	519,334	216,938
Portuguese	458,967	503,854
Spanish	444,696	966,134
Russian	359,677	226,602

ten largest Wikipedias (as of December 2008), along with the number of articles and approximate number of contributors[2].

Due to the ambiguity of certain concepts (e.g. "bank" as "river bank" vs. "financial bank"), Wikipedia adopts the use of unique identifiers to refer to these distinct concepts. Additionally, it also provides *disambiguation pages* which list candidate senses for the target concepts and links to their corresponding articles.

Wikipedia implements a citation framework which allows users to embed citations to other Wikipedia articles that describe and explain the linked concept. For example, let us consider the following Wikipedia snippet (extracted from the article on "natural language processing"):

> **Natural language processing (NLP)** is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. In theory, natural language processing is a very attractive method in the HCI field.

[2] `http://meta.wikimedia.org/wiki/List_of_Wikipedias#Grand_Total`.

> Natural language understanding is sometimes referred to as an AI-complete problem because it seems to require extensive knowledge about the outside world and the ability to manipulate it.

In this snippet, the editors cited the following concepts: "computer science", "linguistics", "HCI", "natural language understanding", and "AI-complete". Each of these *salient* concepts are relevant to the concept of "natural language processing" and by citing and linking them to their respective Wikipedia entries, the reader can navigate to them and gain a better understanding of both the "natural language processing" concept, as well as of its interrelated notions.

Wikipedia provides a very flexible annotation scheme for these citations, which can be linked to an either existing or non-existing article. The latter is called a *stub* and represents a place-holder for a concept which should be formally defined and elaborated on in the future. A citation has a surface form (*anchor*) which can be different from the exact representation of the cited concept (i.e. the cited article title). To illustrate, let us consider the source version of the previously reproduced snippet:

> "Natural language processing" ("NLP") is a field of [[computer science]] and [[linguistics]] concerned with the interactions between computers and human (natural) languages. In theory, natural language processing is a very attractive method in the [[human-computer interaction | HCI]] field. [[Natural language understanding]] is sometimes referred to as an [[AI-complete]] problem because it seems to require extensive knowledge about the outside world and the ability to manipulate it.

The notation [[human-computer interaction | HCI]] indicates a citation to the concept entitled "human-computer interaction;" however, rather than referring to the concept using its original name, the editor opted to use the surface form "HCI" as an anchor text which will be rendered and seen by the reader. Any subsequent clicks on this anchor will guide the reader to the "human-computer interaction" article. In some cases the author is satisfied with the original concept name, as is the case for [[computer science]], where it represent a citation to the exact article entitled "computer science". In either cases the surface form used to cite the concept is referred to as an *anchor*.

Wikipedia also features *interlanguage links*, which are hyperlinks employed by Wikipedia to explicitly connect articles defining the same concept in different languages. For instance, the English article for *bar (unit)* is connected, among others, to the Italian article *bar (unitá di misura)* and the Polish article *bar (jednostka)*. On average, about half of the articles in a Wikipedia version include interlanguage links. Their number varies for each article from an average of five in the English Wikipedia, to ten in the Spanish Wikipedia, and as many as 23 in the Arabic Wikipedia.

CHAPTER 3

RELATED WORK

Models devised for semantic relatedness tasks often address the semantics on the atomic level, reflected in the semantics between single words, as well as on a higher level encountered in the semantics of larger textual constructs should they be sentences, paragraphs, or documents. Each of these levels comes with its own challenges.

On the word level, meaning can be largely ambiguous. For example, the word "crane" could bring to mind "bridge crane" or "overhead crane," a meaning tied to the perceived machine aspect of "crane." This however should be overlooked if we inquire about the relatedness of the word-pair "crane" and "bird," since the existence of "bird" disambiguates the intended meaning of "crane;" hence an automatic system, much like humans, should be able to use this to decode the anticipated meaning. Since the semantic representation on the word level is unaware of the intended meaning, it has to be generic and inclusive of all possible senses of the target word, with limited bias to one sense over the other. This generic representation of meaning makes evaluation on the atomic level a much harder challenge than on the textual level.

At the higher textual abstraction level, many additional clues to the intended meaning can be observed and aggregated in the given textual context to evolve a stricter and unambiguous semantic representation. This representation is more natural as it resembles human adaptation of semantics in everyday life and, for the most part, still relies on the atomic representation of semantics on the word level.

One additional challenge that I identify in semantic relatedness tasks is language, since it plays an important role as a carrier of semantics. Languages have different degrees of linguistic modifiers (inflections) to express different grammatical categories such as tense, mood, voice, aspect, person, number, gender and case. While these modifiers can assist in disambiguating the

intended meaning of text by a reader, they impose an additional degree of difficulty for semantic models to decode and process, due to the high degree of variability in the surface form of words.

This leads to the conclusion that in order to truly study the semantic relatedness task, it needs to be evaluated on both the word and the text level. Additionally, we should study it in the milieu of multiple languages for a better and more robust understanding of the strengths and weaknesses each semantic model entails.

3.1. Word-to-Word Semantic Relatedness

There is a relatively large number of word-to-word similarity metrics that were previously proposed in the literature, ranging from distance-oriented measures computed on semantic networks or taxonomies, to metrics based on models of distributional similarity learned from large text collections. From these, I choose to focus on three corpus-based and six knowledge-based metrics, selected mainly for their observed performance in other natural language processing applications.

3.1.1. Corpus-Based

Corpus-based measures of word semantic similarity try to identify the degree of relatedness between words using information exclusively derived from large corpora. There are three corpus-based measures that have been used more frequently: (1) pointwise mutual information [86], (2) latent semantic analysis [42], and (3) explicit semantic analysis [19].

3.1.1.1. Pointwise Mutual Information

The pointwise mutual information using data collected by information retrieval (PMI-IR) was suggested by Turney [86] as an unsupervised measure for the evaluation of the semantic similarity of words. It is based on word co-occurrence using counts collected over very large corpora (e.g. the Web). Given two words w_1 and w_2, their PMI-IR is measured as:

$$\textit{PMI-IR}(w_1, w_2) = log_2 \frac{p(w_1 \& w_2)}{p(w_1) * p(w_2)} \quad (1)$$

which indicates the degree of statistical dependence between w_1 and w_2, and can be used as a measure of the semantic similarity between w_1 and w_2. From the four different types of queries suggested by Turney [86], the $NEAR$ query (co-occurrence within a ten-word window), represents a balance between accuracy (results obtained on synonymy tests) and efficiency (number of queries to be run against a search engine). Specifically, the following query is used to collect counts from the AltaVista search engine.

$$p_{NEAR}(w_1 \& w_2) \simeq \frac{hits(w_1\ NEAR\ w_2)}{WebSize} \tag{2}$$

Thus, by approximating $p(w_1 \& w_2)$ from Equation 1 with $p_{NEAR}(w_1 \& w_2)$, and estimating $p(w_i)$ to be equal to $hits(w_i)/WebSize$, the following PMI-IR measure is obtained:

$$\textit{PMI-IR}(w_1, w_2) = log_2 \frac{hits(w_1\ AND\ w_2) * WebSize}{hits(w_1) * hits(w_2)}. \tag{3}$$

Since Turney [86] performed evaluations of synonym candidates for one word at a time, the $WebSize$ value was irrelevant in the ranking. In Chklovski & Pantel [12], the $WebSize$ was set to 7x10^{11} in co-occurrence experiments involving Web counts.

3.1.1.2. Latent Semantic Analysis

Another corpus-based measure of semantic similarity is the latent semantic analysis (LSA) proposed by Landauer et al. [42]. In LSA, term co-occurrences in a corpus are captured by means of a dimensionality reduction operated by a singular value decomposition (SVD) on the term-by-document matrix $\mathbf{T}$ representing the corpus.

SVD is a well-known operation in linear algebra, which can be applied to any rectangular matrix in order to find correlations among its rows and columns. In the implementation used for this thesis, SVD decomposes the term-by-document matrix $\mathbf{T}$ into three matrices $\mathbf{T} = \mathbf{U}\mathbf{\Sigma_k}\mathbf{V}^T$ where $\mathbf{\Sigma_k}$ is the diagonal $k \times k$ matrix containing the k singular values of $\mathbf{T}$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k$,

and $\mathbf{U}$ and $\mathbf{V}$ are column-orthogonal matrices. When the three matrices are multiplied together, the original term-by-document matrix is re-composed. Typically k' is chosen such that $k' \ll k$, obtaining the approximation $\mathbf{T} \simeq \mathbf{U}\mathbf{\Sigma}_{\mathbf{k}'}\mathbf{V}^T$.

The first component matrix contains vectors of orthogonal values from the original rows, the second matrix contains vectors of orthogonal values from the original column values, and finally the third matrix is composed of scaling values, such that when they are all multiplied together, we obtain the initial term-document matrix.

The dimensionality reduction using SVD entails the abstraction of meaning by collapsing similar contexts while discounting those that are noisy and irrelevant, hence transforming the real world word-context space into a word-latent-concept space which achieves a much deeper and concrete semantic representation of the words.

In the generalized version of latent semantic analysis [54], rather than constructing a term-document matrix, the model constructs a term-term matrix utilizing pointwise mutual information between the terms found within a context window. The singular value decomposition factorization is then applied to reduce the dimensions of the matrix. This generalized version of LSA was proven to outperform the original LSA [54]. A popular implementation of this generalized LSA method is the Infomap NLP Software package[1], which was used in all the evaluations.

Whether using the original LSA or its generalized version, the relatedness in the resulting vector space is then measured using the standard cosine similarity metric. LSA can be viewed as a way to overcome some of the drawbacks of the standard vector space model (sparseness and high dimensionality). In fact, the LSA similarity is computed in a lower dimensional space, in which second-order relations among terms and texts are exploited. Note also that LSA yields a vector space model that allows for a *homogeneous* representation (and hence comparison) of words, word sets, and texts.

Table 3.1.1.2 shows an example of the LSA semantic vector for the term "automobile." Most of the listed terms are highly relevant, and some of them are close synonyms with the notion of "automobile;" however we can still find some outliers such as "alfa." While most probably the

[1] `http://infomap-nlp.sourceforge.net/`

latter reference alludes to the “Alfa Romeo” automobile brand, the concept of “Alfa Romeo” is not overtly expressed and cannot be inferred easily.

3.1.1.3. Explicit Semantic Analysis

Another corpus-based measure of relatedness that is frequently used is the explicit semantic analysis (ESA) [19], which uses encyclopedic knowledge found in Wikipedia in an information retrieval framework to generate a semantic interpretation of words. ESA relies on the distribution of words inside the encyclopedic descriptions. Since encyclopedic knowledge is typically organized into concepts (or topics), each concept is further described using definitions and examples. ESA takes advantage of this organization by building semantic representations for a given word using a word-concept association, where the concept represents a Wikipedia article. In this vector representation, the semantic interpretation of a word is modelled as a semantic vector consisting of all the concepts (Wikipedia articles) in which the word appears weighted by its occurrence frequency. Furthermore, the semantic interpretation of a text fragment can be modelled as an aggregation of the semantic vectors of its individual words. Such a representation reduces any inherent ambiguity in the text fragment introduced by polysemous terms and promotes context relevant concepts in the feature-space.

In this vector representation, each encyclopedic concept is assigned a weight, calculated as the $tf.idf$ of the given word inside the concept's article. Formally, let C be the set of all the Wikipedia concepts, and let a be any content word, whose ESA concept vector is represented as $\vec{a}$:

$$\vec{a} = \{\langle w_1, c_1\rangle, \langle w_2, c_2\rangle \ldots \langle w_n, c_n\rangle\}, \tag{4}$$

where w_i is the weight of the concept c_i with respect to a. ESA assumes the weight w_i to be the term frequency tf_i of the word a in the article corresponding to concept c_i.

The ESA semantic relatedness between the words in a given word pair is then measured as the cosine similarity between their corresponding vectors [19].

FIGURE 3.1. Explicit semantic analysis framework adopted from [19]

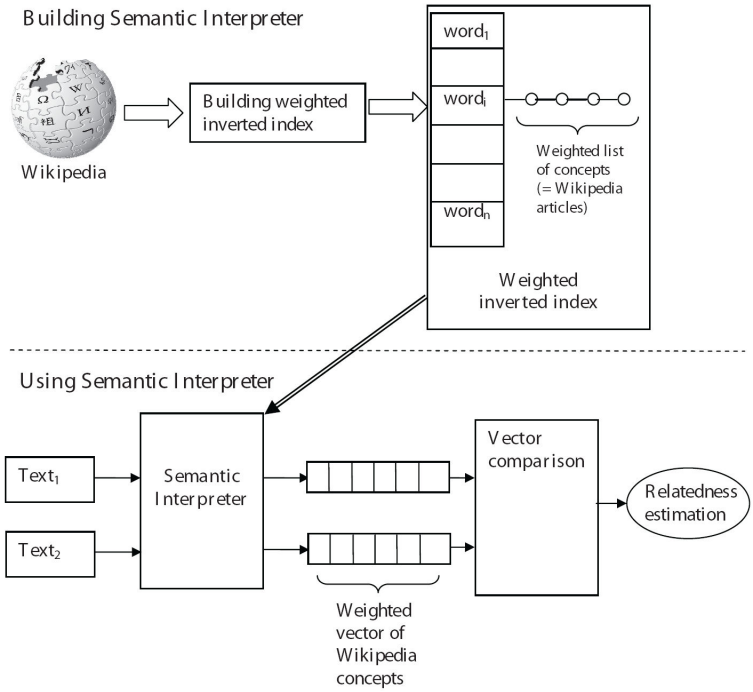


Figure 3.1 demonstrates the explicit semantic analysis framework. As part of it, Wikipedia articles are processed to generate an inverted-index of all the terms in Wikipedia. Given two text fragments, the semantic interpreter generates an ESA context vector for each fragment. The context vectors are then compared to induce a semantic relatedness score. Table 3.1.1.3 shows an example of the ESA semantic vector for the term "automobile." While most of the listed concepts seem to be highly relevant, some of them are not useful in describing the notion of "automobile." For example, while there is no doubt that the concepts "Effects of the automobile on societies" and "Manufacturing industries of Japan" are relevant to the query "automobile," they do not contribute considerably to its meaning.

3.1.2. Knowledge-Based

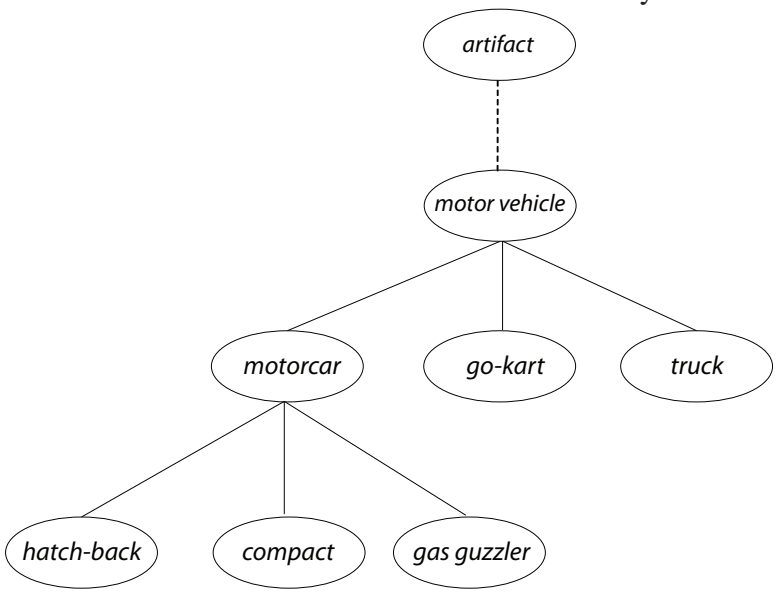
There are a number of measures that were developed to quantify the degree to which two words are semantically related using information drawn from semantic networks [11]. I present below several measures[2] found to work well on the WordNet hierarchy: Leacock & Chodorow

[2]Note that all the word relatedness measures are normalized so that they fall within a 0 to 1 range. The normalization is done by dividing the relatedness score obtained using a given measure by the maximum possible score for that measure.

[44], Lesk [49], Wu & Palmer [89], Resnik [74], Lin [52], and Jiang & Conrath [36]. Additionally, I also introduce one metric based on Roget's Thesaurus [35].

Note that all WordNet metrics are defined between senses, rather than words, but they can be easily turned into a word-to-word similarity metric by selecting for any given pair of words those two meanings that lead to the highest sense-to-sense similarity[3].

FIGURE 3.2. Partial WordNet Hierarchy



The measures below were selected based on their observed performance in other language processing applications, and for their relatively high computational efficiency.

3.1.2.1. Lesk [49]

In this model, the relatedness of two senses is defined as a function of the overlap between the corresponding definitions/glosses, as provided by a dictionary. It is based on an algorithm proposed by $Lesk$ [49] as a solution for word sense disambiguation. The application of the Lesk similarity measure is not limited to semantic networks, and it can be used in conjunction with any dictionary that provides word definitions.

[3]This is similar to the methodology used by McCarthy et al. [55] to find similarities between words and senses starting with a sense-to-sense similarity measure.

3.1.2.2. Leacock and Chodorow [44]

This is one of the simplest knowledge-based models where relatedness is determined as a function of the shortest-path in the WordNet graph as follows:

$$Rel_{lch} = -\log \frac{length}{2 * D} \tag{5}$$

where $length$ is the length of the shortest path between two senses using edge-counting, and D is the maximum depth of the taxonomy. Using this metric, the relatedness between "motor vehicle" and "compact" is $-\log \frac{2}{2*D}$ since the two senses are separated by "motorcar" (Figure 3.2).

3.1.2.3. Wu and Palmer [89]

In this model, the relatedness metric accounts for the depth of two given senses in the WordNet taxonomy as well as the depth of the least common subsumer (LCS) as follows:

$$Rel_{wup} = \frac{2 * depth(LCS)}{depth(sense_a) + depth(sense_b)}. \tag{6}$$

3.1.2.4. Resnik [74]

Unlike the previous metrics which only consider the topology of WordNet, the measure introduced by $Resnik$ incorporates additional statistical information, namely the information content (IC) of the LCS of two senses, as follows:

$$Rel_{res} = IC(LCS) \tag{7}$$

where IC is defined as:

$$IC(c) = -\log P(c) \tag{8}$$

and $P(c)$ is the probability of encountering an instance of sense c in a large corpus.

3.1.2.5. Lin [52]

This model goes one step further by incorporating a normalization factor consisting of the information content of the two input senses to $Resnik$'s measure of similarity.

$$Rel_{lin} = \frac{2 * IC(LCS)}{IC(sense_a) + IC(sense_b)} \tag{9}$$

3.1.2.6. Jiang and Conrath [36]

This model introduces an alternative interpretation of semantic relatedness by discounting the information content of the least common subsumer of $sense_a$ and $sense_b$ from the information contents of the individual senses:

$$Rel_{jnc} = \frac{1}{IC(sense_a) + IC(sense_b) - 2 * IC(LCS)}. \tag{10}$$

3.1.2.7. Roget [35]

Finally, the last relatedness metric considered is $Roget$ [35]. Similar to the previously introduced models, $Roget$ adopts the edge-counting strategy; however it utilizes the 1987 edition of Penguin's Roget's Thesaurus of English Words and Phrases. The relatedness is calculated as the minimum path between two senses in Roget's taxonomy:

$$Rel_{Roget} = D - length(sense_a, sense_b) \tag{11}$$

where D is the maximum distance (16).

3.2. Text-to-Text Semantic Relatedness

Measures of semantic relatedness have traditionally been defined between words or senses, and much less between text segments consisting of two or more words. The emphasis on word-to-word relatedness metrics is probably due to the availability of resources that specifically encode relations between words or senses (e.g. WordNet), and the various testbeds that allow for their evaluation (e.g. TOEFL or SAT analogy/synonymy tests). Moreover, the derivation of a text-to-text measure of relatedness starting with a word-based semantic relatedness metric may not be straightforward, and consequently most of the work in this area has considered mainly applications of the traditional vectorial model, occasionally extended to n-gram language models.

3.2.1. Vectorial Similarity

3.2.1.1. Lexical-based

One of the earliest applications of text relatedness is perhaps the vectorial model in information retrieval, where the most relevant document to an input query is determined by ranking documents in a collection in decreasing order of their relatedness to the given query [80]. Text relatedness has also been used for relevance feedback [75] and text classification [37], word sense disambiguation [49, 82], extractive summarization [51, 81], and automatic evaluation of machine translation [65]. Measures of text relatedness were also found useful for the evaluation of text coherence [43].

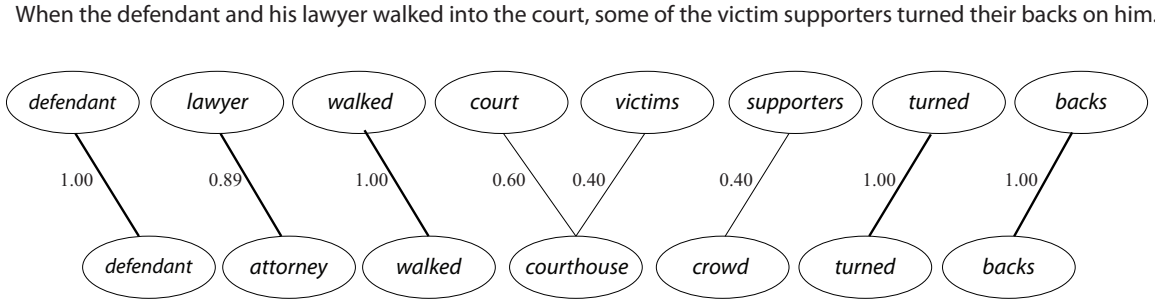
With few exceptions, the typical approach to finding the relatedness between two text segments is to use a simple lexical matching method, and produce a relatedness score based on the number of lexical units that occur in both input segments. Improvements to this simple method have considered stemming, stop-word removal, part-of-speech tagging, longest subsequence matching, as well as various weighting and normalization factors [79]. While successful to a certain degree, these lexical relatedness methods cannot always identify the *semantic* relatedness of texts. For instance, there is an obvious relatedness between the text segments *"we own a pet"* and *"I love animals"*, but most of the lexical-matching text relatedness metrics will fail in identifying any kind of connection between these texts due to vocabulary mismatch.

3.2.1.2. Feature-based

In the feature-based models, each word is expanded using some feature space generating its representative feature vector. Examples of such models are latent semantic analysis [42] and explicit semantic analysis [19] discussed earlier. Since these are vectorial models, the document representation is just a vectorial aggregate of the feature vectors of the terms making up the document. The projection of context from word-space to feature-space leads to a richer representation and greatly reduces (if not eliminates) the vocabulary mismatch problem.

3.2.2. Bipartite Graph Matching

FIGURE 3.3. Bipartite graph matching example



More recently, newly proposed text-to-text relatedness methods [33, 58] (explained in more detail below) utilize a bipartite-graph matching strategy to aggregate word-to-word relatedness among text constituents into one text relatedness score.

The basic premise of these methods lie in the use of an underlying word-to-word relatedness to reach the best semantic alignment among all possible word-pairs (constructed by pairing words from one text fragment with words from the other text fragment).

Formally, let T_a and T_b be two text fragments with vocabulary V_a and V_b, respectively. Let $Rel(u, v)$ be a word-to-word semantic relatedness score for the word-pair (u, v). After removing all non-content pairing terms (e.g. stop-words), the best semantic alignment is the set A, such that:

(12) $$A = \{(t_a, t_b) \mid t_a \in V_a, t_b \in V_b, Rel(t_a, t_b) \geq Rel(t_a, t_i), 1 \leq i \leq |V_b|\}.$$

Once the best alignment is reached (Figure 3.3), the methods differ in how they weight and aggregate these alignments.

In Mihalcea et al. [58], the model takes into account the *specificity* of words, so that a higher weight is given to a semantic matching identified between two specific words (e.g. *collie* and *sheepdog*), and places less importance on the relatedness measured between generic concepts (e.g. *get* and *become*). While the specificity of words is already measured to some extent by their depth in the semantic hierarchy, they are reinforced with a corpus-based measure of word specificity, based on the distributional information learned from large corpora. The best alignments are weighted with the corresponding word specificity, summed up, and normalized by the length of each text segment. The relatedness between the input text segments T_a and T_b is therefore determined using the following scoring function:

(13) $$Rel(T_a, T_b) = \frac{1}{2}\left(\frac{\sum_{t_a \in V_a}(maxRel(t_a, T_b) * idf(t_a))}{\sum_{t_a \in V_a} idf(t_a)} + \frac{\sum_{t_b \in V_b}(maxRel(t_b, T_a) * idf(t_b))}{\sum_{t_b \in V_b} idf(t_b)}\right)$$

Where $maxRel(t_a, T_b)$ is the score of the best alignment of the term $t_a \in V_a$ with its best matching counterpart in V_b. Similarly, $maxRel(t_b, T_a)$ is the score of the best alignment of the term $t_b \in V_b$ with its best matching counterpart in V_a.

In Islam & Inkpen [33] the STS model adopts a corpus-based word-to-word relatedness metric named second-order co-occurrence point-wise mutual information ($SOCPMI$). In this word-to-word relatedness model, each word is expressed as a semantic vector of the words in its immediate context. In its text-to-text generalization, the semantic relatedness score is further augmented with the Longest Common Subsequence (LCS) string matching score to overcome lexical variations before it is aggregated and normalized by the harmonic mean of the text fragments' size.

3.3. Multilingual Semantic Relatedness

Due to the continuous growth of the Internet, and the increasing need for multilingual NLP applications, many NLP tasks need to be carried out in a variety of languages. Most researchers have responded to this desideratum by trying to simplify the problem and restricting it to one linguistic model, most of the time in English, as most of the tools and resources to date have been generated in this language. For example, the authors in [6] were working on a sentiment analysis system called Lydia that would identify positive and negative news for a number of languages. Their approach was to apply machine translation on the text in the foreign languages that required classification, and then perform sentiment analysis in English, and project the labels back onto the original text. Similarly, for the Cross-lingual Information Retrieval Task [40], the most typical used technique was query translation using bilingual dictionaries or machine translation supplemented with English-based information retrieval.

As an alternative to this restrictive mono-lingual representation, some recent research in Natural Language Processing has focused on utilizing multilingual resources to improve performance in domains like sentiment analysis and subjectivity. In Banea et al. [5], the authors have explored the use of parallel multilingual corpora to improve subjectivity classification in a target language. Similarity, authors in [4] have investigated the use of multilingual contexts thus reframing the traditional Word Sense Disambiguation task. By leveraging on the translations of the annotated contexts in multiple languages, they were able to build a multilingual thematic space which better disambiguates target words. In the part-of-speech tagging domain, Cucerzan & Yarowsky [14] utilized an annotated corpora and bilingual dictionaries to discover part-of-speech tags in foreign languages. This has been the practice especially when dealing with languages that have sparse resources; hence by leveraging resources and tools originating from resource rich languages, more robust predictions can be extracted.

Also relevant to this work, authors in [7] examined the notion that the semantic distances between document vectors within a language correlate with the distances observed between their corresponding vectors in a parallel corpora. These findings provide clues about the possibility of reliable semantic knowledge transfer across language boundaries. To my knowledge, there

has been no work in the literature to incorporate corpus-based multilingual features to improve semantic relatedness, hence this research is novel and breaks new boundaries in this domain.

TABLE 3.1. Top 20 concepts from the LSA semantic vector of "automobile"

Weight	Wikipedia Articles
1.00	automobile
0.93	motor
0.92	motorcycle
0.91	car
0.91	automotive
0.90	benz
0.90	chrysler
0.90	lamborghini
0.90	daimler
0.90	truck
0.90	sunbeam
0.89	ford
0.88	alfa
0.88	peugeot
0.88	porsch
0.87	honda
0.87	mercedes
0.87	automaker
0.86	marque
0.86	chevrolet

TABLE 3.2. Top 20 concepts from the ESA semantic vector of "automobile"

Weight	Wikipedia Articles
955.88	History of the automobile
890.73	Karl Benz
834.99	Automobile industry in China
795.40	Daimler-Motoren-Gesellschaft
663.40	Passenger vehicles in the United States
629.97	Korean automobile industry
624.61	Nanjing Automobile (Group) Corporation
573.66	Carl G. Fisher
567.59	Autorack
563.78	Steve Butler
514.57	Automotive engineering
508.33	List of defunct United States automobile manufacturers
494.44	Brass Era car
494.18	Automobile industry in India
485.47	Manufacturing industries of Japan
476.93	Fiat
471.76	MG Rover Group
452.08	Effects of the automobile on societies
431.32	Ferdinand Anton Ernst Porsche

CHAPTER 4

SALIENT SEMANTIC ANALYSIS

4.1. Theoretical Motivation and Considerations

In this chapter, I seek to briefly address the theoretical foundation of this work. Specifically, I present the discourse comprehension problem from a psycholinguistic point of view and refer to the levels of comprehension adopted by mainstream psycholinguists. Additionally, I describe the prevailing model in simulating the comprehension process and address the role of concepts in the semantic interpretation of discourse and how the model relates to the current assumptions. Finally, I introduce a detailed and formal description of the salient semantic analysis framework.

4.1.1. Discourse Comprehension

> Discourse comprehension, from the viewpoint of a computational theory, involves constructing a representation of a discourse upon which various computations can be performed, the outcomes of which are commonly taken as evidence for comprehension. Thus, after comprehending a text, one might reasonably expect to be able to answer questions about it, recall or summarize it, verify statements about it, paraphrase it, and so on.
>
> *Walter Kintsch*

Discourse processing and comprehension is an active field of research in psychology since it opens a window to literally all cognitive functions such as perception, reasoning, problem solving, inference, and memory. At its core, it is concerned with the interpretation and the representation of meaning from written or oral messages, and its integration in our perceived world view/knowledge from the perspective of readers/listeners. This process involves the interaction of multiple levels of representation to reach a coherent understanding of the message. For example, let us consider the following sentence:

> "The wisps of fog over the warm river, along with the dark muddy *banks* and the cloudy night skies lent an ominous feeling."

TABLE 4.1. The definition of the noun *bank*, featuring ten different senses and samples of usage scenario extracted from WordNet

bank (sloping land (especially the slope beside a body of water)) “they pulled the canoe up on the bank”; “he sat on the bank of the river and watched the currents”
depository financial institution, bank, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) “he cashed a check at the bank”; “that bank holds the mortgage on my home”
bank (a long ridge or pile) “a huge bank of earth”
bank (an arrangement of similar objects in a row or in tiers) “he operated a bank of switches”
bank (a supply or stock held in reserve for future use (especially in emergencies))
bank (the funds held by a gambling house or the dealer in some gambling games) “he tried to break the bank at Monte Carlo”
bank, cant, camber (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)
savings bank, coin bank, money box, bank (a container (usually with a slot in the top) for keeping money at home) “the coin bank was empty”
bank, bank building (a building in which the business of banking transacted) “the bank is on the corner of Nassau and Witherspoon”
bank (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)) “the plane went into a steep bank”

We can easily identify the intended meaning of the noun *bank*, yet in the background, there are multiple cognitive processes required to decode its correct sense, given the multitude of meanings it can carry (see Table 4.1.1 for all the senses of the noun bank in WordNet 3.0 [60]). We disambiguate words by utilizing local coherence between the text fragments (“warm river” and “muddy banks”) and bringing in complementary information on the way they relate from our prior world knowledge.

While discourse psycholinguists usually focus their studies on naturally occurring text observed in literature, which allows for the discovery of meanings that are prevalent in culture, they also formulate synthetic stories representing a micro-world to measure certain aspects of discourse comprehension void of the external influences associated with the real-world.

4.1.2. Evaluation Strategies

Researchers employ different strategies to investigate text-comprehension [25]. Some of them are off-line, and they test meaning representation once the comprehension is complete. For example, the tests are administrated when the subjects finish reading a paragraph or an article which involves recall assessment (how much the subject can recall from the original text), summarization exercise (how successful is the subject in summarizing the original text), Q&A assessment (how accurately the subject can respond to propositions about the text using different formats such as true/false, short answers, etc). Alternatively, researchers also adopt on-line strategies that try to asses meaning representation during the comprehension process by automatically collecting and maintaining logs of reading times [21], or gaze duration [38], as well as interruptive scenarios where the researcher periodically questions the subject's understanding during the comprehension process.

4.1.3. Levels of Discourse Comprehension

The majority of discourse psycholinguists embrace Dijk & Kintsch's [87] model of discourse comprehension levels [24], which are briefly introduced below:

- *Surface code*: it deals with the raw textual representation observed in the text and its syntactic structure. It is usually a short-lived memory of words and phrases in the text.
- *Textbase*: it represents logical propositions entailed by the text, which embody all possible meanings of the text fragments.
- *Situational model*: it illustrates the micro-world that the text is manifesting, which incorporates the inferences drawn from the reader's world knowledge, along with the propositions entailed by the text. On that level, text fragments lose their individuality and become integrated into the reader's world knowledge.

- *Communication level*: it represents understanding the writer's/author's intentions and identifying the pragmatic context of the text.
- *Text genre*: it represents the thematic purpose of the text (e.g. narrative, description, persuasion, exposition, jokes, etc).

To further illustrate these comprehension levels, let us consider a fragment from the sentence previously proposed:

"The dark muddy banks and the cloudy night skies lent an ominous feeling."

The raw text serves as the *surface code* which offers the first level of comprehension. In the *textbase* level, the fragment is processed to extract useful propositions, where a proposition represents a state, event, or action and may have a true or false value with respect to a real or imaginary world ([24]). Each proposition requires a predicate, such as a verb, adjective, adverb, or conjunction, and arguments, where an argument represents a functional role such as an agent, object, or location. Thus our fragment is represented at the *textbase* level as:

- $PROP\ 1$: dark ($OBJECT$ = muddy banks)
- $PROP\ 2$: muddy ($OBJECT$ = banks)
- $PROP\ 3$: cloudy ($OBJECT$ = night skies)
- $PROP\ 4$: and ($PROP\ 1$, $PROP\ 3$)
- $PROP\ 5$: ominous ($OBJECT$ = feeling)
- $PROP\ 6$: lent ($AGENT = X$, $OBJECT = PROP\ 5$)

where X in proposition six represents the author. At the *situational model* level, we draw inferences from our world knowledge and previous experiences which support the evidences (propositions) observed in the text. We now understand that "dark banks" and "night cloudy skies" draw a gloomy and eerie scenario. We further infer that these factors represent the direct cause of the "ominous feeling" that the author is mentioning. At the *communication level*, we realize that the intended purpose of the author is to communicate a sense of fear and anxiety. Finally, on the *text genre level*, we clearly identify this text as a narrative, where the author is providing his account of an ominous situation.

Experimental studies have supported the notion that textbase and situational model comprehension can be clearly distinguished. In a study by McNamara et al. [56], two sets of subjects studied a coherent and well written technical text about the biological processes in the heart. One group was unfamiliar with the topic while the other was familiar. The authors administered multiple experiments to the two groups which involved recall scenarios (to ascertain the textbase comprehension level) and inference questions (to glimpse into the situational model comprehension). They found that this coherent text has improved the comprehension on both the textbase as well as the situational model level for the unfamiliar group. However, the familiar group results indicated that using coherent text vaguely improved the recall (textbase comprehension) and decreased or did not affect the inferencing accuracy (situational model comprehension). This implies that a coherent and well written text is a prerequisite for understanding by subjects possessing low-knowledge levels of a given topic, yet surprisingly the same type of exposition is not suitable for the group familiar with the topic. They concluded that in order to engage readers with a high topical familiarity, the text has to have coherence gaps. These would stimulate and provoke the reader into processing the information on a higher level, thus reaching the situational model comprehension where inferences occur, and where incoherences are resolved.

Additionally, experiments performed by Zwaan [92] demonstrate that the surface code and the textbase levels of comprehension are activated and improved when subjects are reading *belle-lettres* genres, such as fiction, poetry and drama, while the situational model comprehension suffers. On the contrary, the situational model greatly improves and the surface code is reduced when the subject is reading news articles or scientific content. This is due to the fact that when reading literature such as poetry, the subject pays more attention to the stylistic representation of the text and the clever use of words, as this body of text is valued for originality and aesthetics; however when reading news or scientific content, the subject tries to integrate this knowledge through some level of inference into the understanding framework he/she already possesses.

Using these clues about the situational model comprehension, it is fair to deduce that most of our world knowledge representation extracted from text is induced primarily through expository and argumentative text types, and only secondarily through narrative text types. In the expository

text type, the contents are aimed at explaining and analyzing facts, while in the argumentative text type the contents focus on representing arguments for or against an issue.

Since semantic relatedness requires inference using one's world knowledge, it is more appropriate, based on these psycholinguistics findings, to rely on expository collections (such as Wikipedia) or argumentative corpora to construct and simulate world-knowledge representation for machines, much like the way humans infer world-knowledge using the situational model comprehension. It is important to note that while expository collections would entail rich world-knowledge, some relationships tend to be prevalent in the narrative text type (e.g. *picnic-sunshine*, *romance-flower*). Therefore, a complete model should incorporate some supplemental narrative text collections in addition to the expository text for a better coverage.

4.1.4. Construction-Integration: Discourse Comprehension Model

While there are many models to explain and simulate discourse comprehension, one model stands out as the most accurate to date: the construction-integration by Kintsch [39]. In Kintsch [39], knowledge is represented as an associative net where nodes represent concepts or propositions and the edges represent the strength of association between these nodes. Kintsch's model simulates discourse-comprehension in two major phases, a construction and then an integration phase. In the construction phase, the subject constructs all possible meanings of the intended text which directly relate to how ambiguous the text is. These meanings are encoded in a set of propositions. Let us consider the following sentence: "I shot an elephant in my pajamas". This sentence generates multiple propositions based on potential meanings. Its first meaning could be that the subject shot an elephant while wearing pajamas, a second meaning is that the elephant was literally wearing the subject's pajamas when it was shot, while a third meaning is that the elephant was somehow fitting in the pocket of the pajama when the subject shot it. Similarly, in the bank example previously illustrated, there are numerous potential meanings (see Table 4.1.1). Using inferences derived from our world knowledge, we are able to connect these propositions into an associative semantic net, which represents the working memory. These connections indicate the strength of association between the propositions as well as their association with our current knowledge. These propositions are later constrained in the integration phase by activating the correct meaning that preserves the

global coherence of the text. This activation takes into account the contextual clues observed in the contents, in order to disambiguate the intended meaning and filter out irrelevant propositions which do not fit within the overall context.

As there is an understanding that mental data comes in pieces or units, Anderson [3] articulates three different types of units, namely concepts, propositions, and schemata. Concepts are primitives of cognitive representation; the meaning of a concept is attached to a node expressing this concept in a semantic network. Propositions represent constraints on concepts and characterize the smallest unit of meaning that asserts things about the situational model. Schemata signifies a set of related propositions. These three units typically correspond to words, sentences, and passages, respectively. Additionally, Anderson [3] also hypothesizes that propositions and concepts can be treated alike, a notion also embraced by Kintsch [39] in modeling the discourse-comprehension process. This notion is also leveraged by Csomai [13], who approximates Kintsch's model by limiting the semantic network to only keywords (concepts) and ignoring propositions, and applies the ensuing model to the keyword extraction task.

Similarly, in the proposed model I seek to build upon this notion of concept. In my interpretation, I assume that there are key unambiguous concepts that are salient in the context. These concepts are easily and quickly integrated in our working memory during the discourse comprehension process, and allow, in part, for the easy anchoring and recall of information. This is also in line with observations from the field of psycholinguistics, where Graesser et al. [24] noted that accessing familiar world knowledge in the working memory is inexpensive. By using these salient concepts to index the surrounding context, a better concept-based representation of content can be generated that corresponds to some extent with our mental representation.

4.2. Salient Semantic Analysis Model

The model is formulated based on the assumption that a casual reader's representation of a given text entails a mental frame that retains and pairs unambiguously defined concepts observed in text. This articulation allows for easy anchoring and association of these concepts with other relevant mental knowledge maintained by the reader, subsequently leading to easier recall. To validate this assumption, I perform an annotation experiment where two human annotators are

provided with a text containing three paragraphs representing the unformatted and untitled abstract section of Wikipedia about "car"[1]. After reading the abstract, the annotators are asked to recall as many keywords as they can remember from the text. The findings are that 100% of the recalled keywords - listed in Table 4.2 - refer to Wikipedia concepts that are salient in the context. This supports the claim that these salient keywords are activated in the mental frame of the reader, which allows for quick and easy anchoring and recall.

TABLE 4.2. Keywords recalled by the two subjects when presented with the Wikipedia article about "car;" bold faced keywords represent the overlap between subjects

Keywords	
Car	*Truck*
Automobile	*Engine*
Passenger	*India*
$4-Wheels$	*China*
Rail	*Coal*
Fuel	*Locomotive*
Road	*motorcar*

Consequently, semantic profiles are derived based on the Wikipedia corpus by using one of its most important properties – the linking of concepts within articles. The links available between Wikipedia articles, obtained either through manual annotation by the Wikipedia users or using an automatic annotation process, allow for determining the meaning and the saliency of a large number of words and phrases inside this corpus. These links are regarded as clues or salient features within the text, that help define and disambiguate its context. The semantic relatedness of words can be measured by using their concept-based profiles, where a profile is constructed using the co-occurring salient concepts found within a given window size in a very large corpus.

[1] `http://en.wikipedia.org/wiki/Car`

To illustrate, let us consider the following paragraph extracted from a Wikipedia article[2]:

An automobile, motor car or car is a wheeled motor vehicle used for transporting passengers, which also carries its own engine or motor. Most definitions of the term specify that automobiles are designed to run primarily on roads, to have seating for one to eight people, to typically have four wheels, and to be constructed principally for the transport of people rather than goods.

All the underlined words and phrases represent linked concepts, which are disambiguated and connected to the correct Wikipedia article. Therefore, each term in this example can be semantically interpreted as a vector of its neighbouring linked concepts (as opposed to simple words, as done in the other corpus-based measures). For example the word "motor" can be represented as a weighted vector of the salient concepts "automobile," "motor car," "car," "wheel," "motor vehicle," "transport," and "passenger."

In this interpretation, a word is defined by a set of concepts which share its context and they are weighted by their pointwise mutual information.

The method proposed here consists of two main steps. In the first step, starting with Wikipedia, a corpus where concepts and saliency are explicitly annotated is created. Next, this corpus is used to build salient concept-based word profiles, which are utilized to measure the semantic relatedness of words and texts.

4.2.1. Constructing a Corpus Annotated for Concepts and Saliency

A large annotated corpus is created from Wikipedia, by linking salient words and phrases to their corresponding articles.

First, I use the manual links as provided by the Wikipedia users. These links have two important properties that are relevant to the method. On one hand, they represent concepts that are salient for a given context, since according to the Wikipedia guidelines, only those words or phrases that are important to the understanding of a certain text should be linked. On the other hand, the links connect surface forms to Wikipedia articles, thereby disambiguating the corresponding words or phrases. For instance, even if the word "car" is ambiguous, a manual link connecting this word

[2] `http://en.wikipedia.org/wiki/Car`

to the Wikipedia article "motor car" will eventually indicate that the intended meaning is that of "automobile" rather than "rail car."

Next, I use the one sense per discourse heuristic [20], according to which several occurrences of the same word within a discourse tend to share the same meaning. In this case, each additional occurrence of a word or phrase that matches a previously seen linked concept inside a given page is also linked to the same Wikipedia article. Moreover, since the already linked concepts are assumed to be salient for the given text, this property is transferred to the newly linked words or phrases. For example the second occurrence of the word "automobile" in the example shown in Figure 4.2 is also disambiguated and linked to a Wikipedia article even though it was not initially linked by a Wikipedia user. Additionally, since the first occurrence of "automobile" was considered to be salient for this particular text (because of the first link), I assume that the second occurrence will also have this property.

Finally, I use a disambiguation method similar to the one used in the Wikify! system [59], which assigns Wikipedia articles to words or phrases that have a high hyperlinkability (or keyphraseness). Very briefly, this method first determines the phrases that have a high probability (≥ 0.5) to be selected as a keyphrase, which corresponds to a high saliency. This probability is calculated as the number of times a word or phrase appears inside a manually generated link divided by the total number of times that word or phrase appears in Wikipedia (hyperlinked or not). From this set, the words or phrases that have a probability of 95% or higher to point to only one article are tagged with the corresponding article. This disambiguation method can be interpreted as a strengthened most frequent sense heuristic.

Table 4.2.1 shows the collected corpus stats for English, Spanish, Arabic, and Romanian, including the number of pages, the number of links identified by each of the three methods mentioned earlier, and the various thresholds considered for the *most frequent sense* method. The latter are explained in more detail in Section 5.4.

4.2.2. Salient Concept Based Word Profiles

The corpus is processed to generate semantic profiles for words using their most contextually relevant concepts, namely the surface forms linked to Wikipedia articles.

TABLE 4.3. Corpus Statistics for the Wikipedia versions in English, Spanish, Arabic and Romanian

Stats	English	Spanish	Arabic	Romanian
Wikipedia	2009	2009	2011	2011
Number of Pages	4,406,717	477,930	381,089	296,770
Manually Disambiguated	38,973,005	13,331,757	2,622,596	2,803,814
One Sense Per Discourse	29,062,299	10,732,678	1,704,645	2,146,831
Most Frequent Sense	11,237,128	2,228,088	407,214	449,387
Key Phraseness	0.50	0.40	0.20	0.40
Monosemous Probability	0.95	0.85	0.65	0.85

Formally, given a corpus C with m tokens, vocabulary size N, and concept size W (number of unique Wikipedia concepts), a co-occurrence $N \times W$ matrix (E) is generated representing the cumulative co-occurrence frequencies of each of the corpus terms with respect to its contextual concepts (defined by a context window of size k).

$$E_{ij} = f^k(w_i, c_j) \tag{14}$$

where E_{ij} represents the element found at the intersection of the ith row with the jth column in matrix E. f^k is the number of times the term w_i and concept c_j co-occur within a window of k words in the entire corpus. The matrix is further processed to generate a $N \times W$ PMI matrix P:

$$P_{ij} = log_2 \frac{f^k(w_i, c_j) \times m}{f^C(w_i) \times f^C(c_j)} \tag{15}$$

where P_{ij} denotes the element found at the intersection of the ith row and the jth column in the matrix P. $f^C(w_i)$ and $f^C(c_j)$ are the corpus frequencies for the term w_i and concept c_j, respectively.

Each row P_i is further filtered to eliminate irrelevant associations by only keeping the top β_i cells [31] and zeroing the rest. This corresponds to selecting the β highest scoring PMI terms associated with a given row:

$$\beta_i = (log_{10}(f^C(w_i)))^2 \times \frac{log_2(n)}{\delta}, \delta \geq 1 \tag{16}$$

where δ is a constant that is adjusted based on the size of the chosen corpus.

The semantic relatedness involves estimating the strength of the semantic bond between textual entities, be they words or texts. The word-to-word scenario entails evaluating the relatedness of a word pair (e.g. car - $automobile$). On the other hand, the text-to-text scenario involves a text pair, where the text can be a sentence, paragraph, or document.

4.2.2.1. Word Relatedness

To calculate the semantic relatedness of a given word pair, the overlap between the semantic profiles of the words in the word-pair is aggregated to produce a relatedness score. Thus, given the constructed matrix E, I adopt a modified cosine-metric illustrated in Equation 17.

$$Score_{cos}(A, B) = \frac{\sum_{y=1}^{N}(P_{iy} * P_{jy})^{\gamma}}{\sqrt{\sum_{y=1}^{N} P_{iy}^{2\gamma} * \sum_{y=1}^{N} P_{jy}^{2\gamma}}}, \tag{17}$$

The γ parameter allows for the control the weight bias. Additionally, since cosine is a normalized metric that scores one for identical terms, it is negatively impacted by a sparse space as it tends to provide low scores for near synonyms. This creates a large semantic gap between matching terms and strongly related terms. To close this gap and provide more meaningful scores, I also include a normalization factor λ, as shown in equation 18.

(18) $$Sim(A,B) = \begin{cases} 1 & Score_{cos}(A,B) > \lambda \\ Score_{cos}(A,B)/\lambda & Score_{cos}(A,B) \leq \lambda \end{cases}$$

Table 4.2.2.1 shows an example of the SSA semantic vector for the term "automobile." Most of the listed terms are highly relevant and some of them are close synonyms with the notion of "automobile." One interesting observation is that these top salient features encompass multiple senses of automobile (e.g. automobile magazines, automobile designers) which are overlooked notions from the set of top features proposed by LSA and not strongly emphasized in the top features retrieved by ESA.

4.2.2.2. Text Relatedness

To calculate the semantic relatedness between two text fragments, I use the same word profiles built from salient encyclopedic concepts, coupled with a simplified version of the bipartite-graph matching technique proposed in Mihalcea et al. [58] and Islam & Inkpen [33].

Formally, let T_a and T_b be two text fragments of size a and b respectively. After removing all stop-words, the number of shared terms (ω) between T_a and T_b is determined. Then, the semantic relatedness of all possible pairings between non-shared terms in T_a and T_b is calculated, using the normalized form of cosine similarity described previously (Equation 17). These possible combinations are further filtered by creating a list φ which holds the strongest semantic pairings between the fragments' terms, such that each term can only belong to one and only one pair.

Thus, the semantic relatedness between the two text fragments is defined as

(19) $$Sim(T_a, T_b) = \frac{(\omega + \sum_{i=1}^{|\varphi|} \varphi_i) \times (2ab)}{a+b}$$

where ω is the number of shared terms between the text fragments and φ_i is the similarity score for the ith pairing.

TABLE 4.4. Top 20 salient concepts from the SSA semantic vector of "automobile"

Weight	Wikipedia Articles
804	2000s automobiles
594	1990s automobiles
144	1930s automobiles
120	1920s automobiles
105	1940s automobiles
104	1900s automobiles
96	1910s automobiles
91	Automobile magazines
77	Automobile engines
62	Automobile designers
46	Automobile museums in the United States
38	Automobile maintenance
36	First automobile made by manufacturer
35	Automobile awards
29	Automobile history eras
23	Automobile layouts
23	Automobile transmissions
23	One-off automobiles
22	1890s automobiles

4.2.3. Multilinguality in Semantics

I further examine the abstraction of semantics from the choice of the underlying language by exploring whether meaning can be successfully carried across many languages with minimal

dilution. Thus, I seek to aggregate multiple monolingual models so that similarity is not language dependent anymore, but it pervades language boundaries. Since I am focusing on both the word-to-word and text-to-text similarity, each of them will be discussed independently.

4.2.3.1. Word-to-Word

For a given word pair $P =< w_a, w_b >$ formulated in a source language, its relatedness score is computed based on the initial monolingual (source) model. Then, each pair (as a unit) is translated in another language (target language). This way a decision on word choice is influenced by both w_a and w_b, as well as by the relatedness they pose. As experiments involve a set of target languages, the relatedness score obtained for the pair in each one of them is aggregated with the source language score using simple averaging. Thus, the pair P will be defined by a score computed over a number of monolingual models, and receive reinforcement from other languages, when dealing with highly ambiguous terms.

4.2.3.2. Text-to-Text

In the case of text-to-text relatedness, a similar path to the one proposed in the text-to-text monolingual model from Section 4.2.2.2 is pursued. For two text fragments, T_a and T_b in a source language, their translation is obtained by processing each fragment independently. This is motivated by the fact that sufficient context is available in each fragment to allow for its disambiguation in order to provide an accurate translation. Upon computing the relatedness score based on a monolingual framework in both the source and the target languages, the scores are aggregated by averaging them over all the monolingual models.

Thus, the relatedness between the two contexts is strengthened with every additional language. As salient concepts permeate from every language, they lend their disambiguating power to every monolingual fragment, thus allowing for a clearer relatedness relationship to transpire. This relationship is not language dependent anymore, as it manages to capture the quintessential meaning of the fragments in question. To provide the reader with a better understanding, I will

make a parallel with Plato's Allegory of the Cave [68], where a group of prisoners are chained in such a way so that from early childhood they are continuously facing a wall, on which shadows of various objects are cast by the fire behind them. Since they have never been able to turn their head and see anything else but the images before them, they believe that the shadows are the *real* objects, and cannot grasp the reality of the situation. Plato calls the actual objects *forms* or *ideas*, as once they are grasped, they impart us with the ability to recognize *mimes*, corruptions of perfect forms, or, in the allegory, the shadows on the cave's wall. Returning to the multilingual relatedness discussion, by allowing us to access the monolingual spaces, or the *mimes* in our allegory, and aggregate them through superpositions into a more clearly defined *form* or *idea*, that is disentangled from the mere "shadows," we can aspire to reach the true "idea" of semantic relatedness. Therefore semantic relatedness is not an intra-language concept anymore, as text fragments are able to take various forms based on the language of choice, yet the ultimate relationship between the fragments is not only maintained, but augmented, and purified from noise, once abstraction is made of the underlying language.

CHAPTER 5

EXPERIMENTAL SETUP

In this chapter, I introduce the datasets that have been utilized in the evaluations. I further describe the SSA model generation framework by focusing on implementation details. Then I propose an evaluation schema and discuss the issues pertaining to the evaluation metrics. I finalize this chapter with a presentation of the strategies adopted for parameter tuning.

5.1. Experimental Data

Evaluations of semantic relatedness in the literature mostly revolve around using synthetic data handpicked by a human expert to test various semantic relatedness levels (e.g. synonymy, near-synonymy, antonymy, collocation, association, etc.), as well as irrelevant or meaningless associations (e.g. $noon-string$). The relatedness level is determined by multiple human judges who assign a relatedness score to a given word/text pair, which is then averaged to eliminate any subjective annotations. Beside employing synthetic data, some evaluations are performed on real-life datasets such as paraphrase detection and short answer grading.

In order to consider a diverse set of testing scenarios, datasets originally developed for English (Section 5.1.1), as well as several multilingual manually constructed datasets that follow both the structure as well as the annotation guidelines proposed by the English data (Section 5.1.2) are used. These datasets are further organized into word-to-word and text-to-text, based on the type of entities participating in the pairing, whether they are words (e.g. car - $automobile$) or text fragments (*answer key - student response*).

5.1.1. English

5.1.1.1. Word Relatedness

To evaluate the effectiveness of the SSA model on word-to-word relatedness, I use three standard datasets that have been widely employed in the literature, namely *Rubenstein and Goodenough*, *Miller-Charles*, and *WordSimilarity-353*, which are described in more detail below.

Rubenstein and Goodenough [76] consists of 65 word pairs ranging from synonymy pairs (e.g., car - $automobile$) to completely unrelated words (e.g., $noon$ - $string$). The participating terms in all the pairs are non-technical nouns annotated by 51 human judges on a scale from 0 to 4, where 0 represents complete unrelatedness, while 4 denotes perfect synonymy.

Miller-Charles [61] is a subset of the Rubenstein and Goodenough dataset, consisting of 30 word pairs. The relatedness of each word pair was rated by 38 human subjects, using a scale from 0 to 4.

WordSimilarity-353 [18], also known as Finkelstein-353, consists of 353 word pairs annotated by 13 human experts, on a scale from 0 (unrelated) to 10 (very closely related or identical). The Miller-Charles set is a subset in the WordSimilarity-353 dataset. Unlike the Miller-Charles data set, which consists only of single generic words, the WordSimilarity-353 set also includes phrases (e.g., *"Wednesday news"*), proper names and technical terms, therefore posing an additional degree of difficulty for any relatedness metric.

5.1.1.2. Text Relatedness

For the text-to-text relatedness task I employ four datasets that have been frequently used in related work. These are *Lee50*, *Li30*, *AG400*, and the *Microsoft Paraphrase Corpus*, and they are expounded on below.

Lee50 [45] is a compilation of 50 documents collected from the Australian Broadcasting Corporation's news mail service. Each document is scored by ten annotators on a scale from 1 (unrelated) to 5 (alike) based on its semantic relatedness to all the other documents. The users' annotation is

then averaged per document pair, resulting in 2,500 document pairs annotated with their similarity scores. Since it was found that there was no significant difference between annotations given a different order of the documents in a pair [45], the evaluations are carried out on only 1225 document pairs after ignoring duplicates.

Li30 [50] is a sentence pair similarity dataset obtained by replacing each of the Rubenstein and Goodenough word-pairs [76] with their respective definitions extracted from the Collins Cobuild dictionary [84]. Each sentence pair was scored by 32 native English speakers using a range from 0 to 4, where 0 denotes completely unassociated texts, while 4 characterizes interchangeable definitions. The annotations were then averaged to provide a single relatedness score per sentence-pair. Due to the resulted skew in the scores toward low similarity sentence-pairs, they selected a subset of 30 sentences from the 65 sentence pairs to maintain an even distribution across the similarity range [50].

AG400 [64] is a domain specific dataset from the field of computer science, used to evaluate the application of semantic relatedness measures to real world applications such as short answer grading. The original dataset consists of 630 student answers along with the corresponding questions and correct instructor answers. Each student answer was graded by two judges on a scale from 0 to 5, where 0 means completely wrong and 5 represents a perfect answer. The correlation between human judges was measured at 0.64. Since the dataset exhibited a large skew in the grade distribution toward the high end of the grading scale (over 45% of the answers are scored 5 out of 5), I followed Li et al. [50] and randomly eliminated 230 of the highest grade answers in order to produce more normally distributed scores and hence calculate a meaningful Pearson correlation.

Microsoft paraphrase corpus (MSR) [16] contains 4076 training ($MSRB$) and 1725 test ($MSRS$) text pairs. Each text-pair is annotated in a binary fashion indicating whether the paragraphs in the text pair are a paraphrase of each other or not. The dataset was compiled from on line news sources and annotated by two annotators. The resulting inter-annotator agreement is 0.83, which serves as an upper bound for the paraphrase detection task. While paraphrase detection is a complex task that might require a higher level of abstraction and understanding, semantic relatedness can serve

as a solid starting point. The dataset was utilized for the text-to-text semantic relatedness task in Mihalcea et al. [58] and Islam & Inkpen [32].

5.1.2. Multilingual

Due to the lack of multilingual semantic relatedness datasets, I construct equivalent sets in the target languages using the same guidelines adopted for the generation and the annotation of their original English counterparts. This involved recruiting native speakers of different languages to serve as human judges and validating the outcome by observing high levels of inter-annotator agreement.

5.1.2.1. Word Relatedness

I build several multilingual datasets based on the standard Miller-Charles [61] and WordSimilarity-353 [18] English word relatedness datasets. To construct the datasets, native speakers of Spanish, Romanian and Arabic, who were also highly proficient in English, were asked to translate the entries in both data sets. The annotators were presented with one word pair at a time, and asked to provide the appropriate translation for each word while taking into account the relatedness expressed within the word pair. The relatedness was meant as a hint to disambiguate the words, when multiple translations were possible.

The annotators were instructed not to use multi-word expressions in their translations. They were also allowed to use replacement words to overcome slang or culturally-biased terms. For example, in the case of the word pair *dollar-buck*, the Arabic annotators were allowed to use دينَار [1] as a translation for "buck". Such substitutions maintain the semantic relations within the word pairs, while at the same time allowing the words to be translated.

To test the ability of the bilingual judges to provide correct translations by using these annotation guidelines, the following experiment was carried. Spanish translations were collected from five different human judges, and then were merged into a single selection based on the annotators'

[1] Arabic for dinars – the commonly used currency in the Middle East.

TABLE 5.1. Manual translation examples that are part of the multilingual *word-to-word* relatedness dataset that leverages the content and structure of *WordSimilarity-353*, originally developed in English

	Word pair		
English	coast - shore	car - automobile	brother - monk
Spanish	costa - orilla	coche - automovil	hermano - monje
Arabic	شَاطيء - سَاحل	عربه - سيَارة	رَاهب - شقيق
Romanian	ţărm - mal	maşină - automobil	frate - călugăr

TABLE 5.2. Manual translation examples that are part of the multilingual *text-to-text* relatedness dataset that leverages the content of $Li30$, originally developed in English

	text pair	
English	The coast is an area of land that is next to the sea.	The shores or shore of a sea, lake or wide river is the land along the edge of it.
Spanish	La costa es un área de terreno que está junto al mar.	Las costas o costa de un mar, lago o extenso río es la tierra a lo largo del borde de estos.
Arabic	السَاحل هو مسَاحة من الرض التي تقع بجوَار البحر	طول حَافة البحر أو إلبحيرَات أو الءنهَار
Romanian	Coasta este o zonă de teren care se află lângă mare.	Ţărmul sau malul unei mări, lac sau fluviu este pământul de-a lungul marginii acestora.

translation agreement; the merge was done by a sixth human judge, who also played the role of adjudicator when no agreement was reached between the initial annotators.

For the translations provided by the five human judges, in more than 74% of the cases at least three human judges agreed on the same translation for a word pair. When the judges did not provide identical translations, they typically used a close synonym. The high agreement between

their translations indicates that the annotation settings were effective in pinpointing the correct translation for each word, even in the case of ambiguous words.

To test the abstraction of semantics from the choice of the underlying language, five additional human experts re-scored the newly constructed Spanish dataset by using the same scale that was used in the construction of the English one. The correlation between the relatedness scores assigned during this experiment and the scores assigned in the original English experiment was 0.86, indicating that the translations provided by the bilingual judges were correct and preserved the semantics of the original word-pair.

Given the validation of the annotation settings obtained for the Spanish dataset, for Romanian and Arabic only one human annotator was used to collect the translations. Table 5.1.2.1 shows examples of translations in the three languages for three word pairs appearing in the data sets.

5.1.2.2. Text Relatedness

I further construct a multilingual text-to-text relatedness dataset based on on the standard $Li30$ [50] corpus originally developed in English.

Native speakers of Spanish, Romanian and Arabic, who were also highly proficient in English, were asked to translate the entries drawn from the English collection. The annotators were presented with one sentence at a time, and asked to provide the appropriate translation into their natal tongue. Since five Spanish and two Arabic annotators were found to translate the English dataset, an arbitrator (native to the language) was charged with merging the candidate translations by proposing one sentence per language.

Furthermore, to test the abstraction of semantics from the choice of underlying language, the annotation experiment carried out in the multilingual word-to-word dataset (Section 5.1.2.1) was repeated. Three different Spanish human experts were asked to re-score the Spanish text-pair translations on the same scale used in the construction of the English collection. The correlation

between the relatedness scores assigned during this experiment and the scores assigned in the original English experiment was $0.77 - 0.86$, indicating that the translations provided by the bilingual judges were correct and preserved the semantics of the original English text-pair.

5.2. SSA Model Generation Framework

I introduce the overview of the SSA model generation framework by illustrating the multi-phase process in Figure 5.1. To enrich this basic structure, I then cover the particulars related to each stage in detail.

5.2.1. Wikipedia Parser

The framework starts with processing Wikipedia raw data; to that end I designed a Wikipedia parser which performs multiple scans on Wikipedia to extract the following:

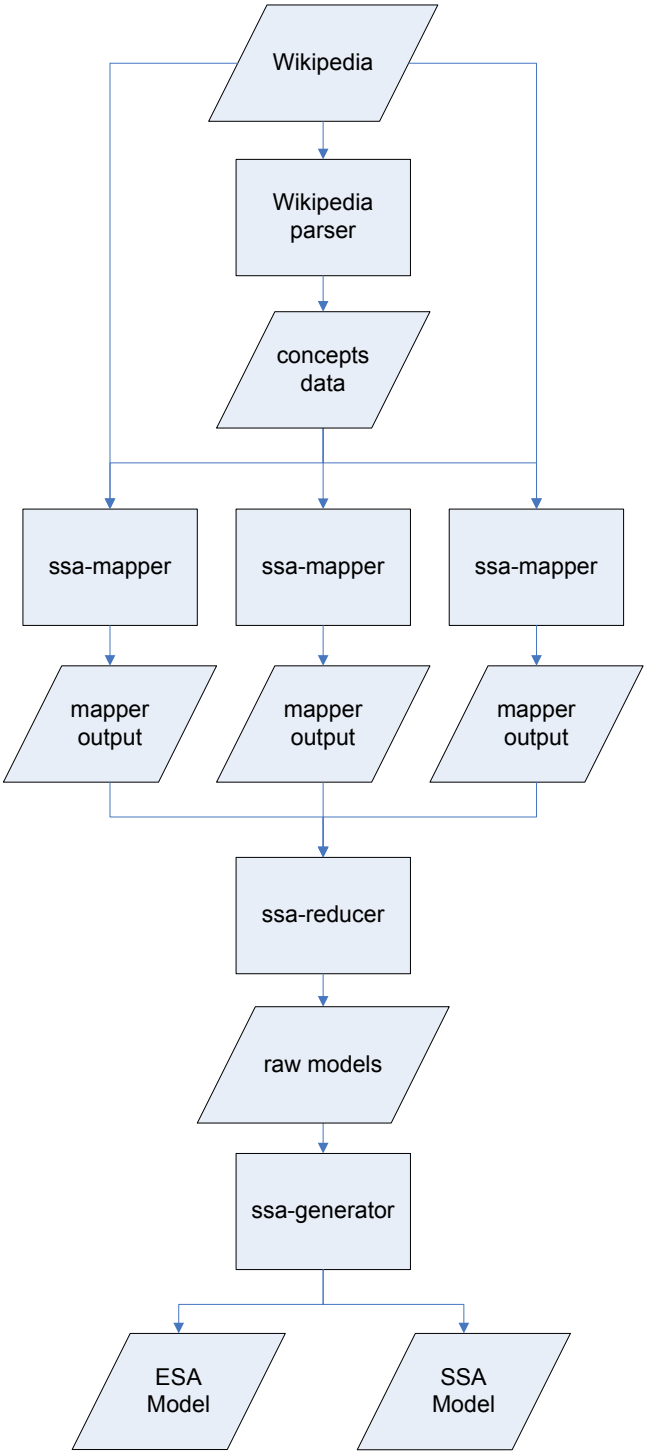
- *Page titles*: titles associated with Wikipedia concepts.
- *Page redirects*: redirect pages which map multiple Wikipedia concepts into one unique concept. For example, the Wikipedia concept "car" is just a redirect pointing to the concept "automobile." This information is valuable in finding alternative representations for concepts.
- *Anchor - concept associations*: since Wikipedia annotations (described in Section 2.4) allow multiple surface forms for concepts during in-article citation, the parser collects these different representations along with their frequencies. Let us consider the following article's source version example extracted from Wikipedia, which showcases the surface form "Britain" that links to the article entitled "United Kingdom:"

  > "In [[United Kingdom|Britain]], there had been several attempts to build steam cars with varying degrees of success."

  This sentence is rendered to Wikipedia visitors as:

  > "In Britain, there had been several attempts to build steam cars with varying degrees of success."

FIGURE 5.1. Salient semantic analysis generation framework



The parser is able to identify that the anchor "Britain" is just another representation (synonym) for the concept "United Kingdom." It also collects how many times this anchor ("Britain") has been used to refer to the "United Kingdom" concept.

- *Anchor - occurrences*: the frequency of anchors in the unformatted Wikipeda text. Based on the previous example, this metric helps keep track of how many times the expression "Britain" appeared in Wikipedia as text. This information paired with the other occurrence metric allows for the application of the wikification step described in Section 4.2.1.

5.2.2. SSA-Mapper

Once the initial parsing of Wikipedia is complete, the outcome of the parsing phase along with the Wikipedia articles are provided to the *ssa-mapper*. In this phase, the article contents are processed to identify (in place) the salient concepts, whether they are annotated by the users, by using one sense per discourse heuristics, or by using the wikification process. Then, the context surrounding each of the identified concepts is indexed by these concepts.

Since this process is parallelizable, multiple *ssa-mappers* were initiated, thus allowing for exectution in a multi-core system or across multiple machines (e.g. using the Hadoop framework[2]).

5.2.3. SSA-Reducer

The output of the mappers is then merged and reduced using the *ssa-reducer* to produce raw co-occurrence models.

5.2.4. SSA-Generator

Finally, the *ssa-generator* takes these raw co-occurrence models along with the filtering criteria and the intended weighting schema (pmi for SSA and $tf.idf$ for ESA) to generate the finalized SSA model along with the ESA model. The same process was intentionally used to produce both models to guarantee fair treatment (e.g. preprocessing, stemming, filtering, etc). The output format of the generated model is a gdbm database[3]. Since gdbm is architecture dependent, the model is also serialized as a flat file for portability across platforms. The gdbm format is a GNU implementation of the standard UNIX dbm library which is supported under many languages (C++, Java, Perl, etc). It implements a file system-based hash table which allows speedy access, while also featuring a negligible memory footprint. This solution offers a good compromise between speed (in memory operations) and space (memory footprint).

[2] `http://hadoop.apache.org`
[3] `http://www.gnu.org/software/gdbm/`

5.3. Evaluation Metrics

When it comes to evaluating relatedness measures, the literature is split on the correct correlation to use. While a number of previous projects adopted the Pearson correlation metric r [31, 35, 64, 85], there are several others that employed the Spearman correlation ρ [19, 30, 66, 91].

Most recently, to address this issue, several publications reported relatedness results using both metrics [26, 63]. While Pearson (parametric with values between -1 and 1) erects constraints on a correlation by imposing a linear relation between the two sets of normally distributed values being compared, Spearman (non-parametric) relaxes this condition and only constrains the consistency of the relative ranking across the two inputs while imposing no distributional requirements.

TABLE 5.3. An example of Pearson versus Spearman correlations

Word pair	H	A	B
$car - automobile$	3.92	1×10^{-1}	2.95
$tool - implement$	2.95	1×10^{-3}	3.92
$journey - car$	1.16	1×10^{-5}	0.08
$noon - string$	0.08	1×10^{-7}	1.16
r	1.00	0.74	0.77
ρ	1.00	1.00	0.60

To better demonstrate the strengths and weaknesses of each metric, let us consider the example illustrated in Table 5.3. The table presents four word pairs selected from the Miller-Charles dataset, along with their human assigned relatedness scores. A and B are two hypothetical systems that score these pairs as shown in the table. System A maintains the exact order of the ideal scores reporting a perfect Spearman correlation, however, due to the inability of the system to correctly quantify the weight of the semantic bonds between the words in each word pair, the Pearson correlation results in a modest 0.74. On the other hand, system B scores were produced by swapping the gold standard scores of the first two pairs as well as the last two pairs. This results

in poor ranking performance, as illustrated by the low Spearman score (0.60). However, Pearson is more resilient in this scenario, resulting in a correlation of 0.77. This is largely due to the fact that the absolute difference between two incorrect values and their gold standard values is relatively small, hence the penalty imposed on the correlation is proportional.

I believe that both metrics are important for the evaluation of semantic relatedness, since an ideal system should be able to demonstrate a linear relationship with the human judgment and also to satisfy the relative order imposed by these judgments. In a sense, a good system should maintain the correct ranking between word/text pairs, and at the same time correctly quantify the strength of the relatedness for a given word/text pair. I am therefore reporting both correlation metrics, as well as the harmonic mean of the Pearson and Spearman metrics μ, which penalizes the system if it fails to simultaneously achieve these two goals.

$$\mu = \frac{2r\rho}{r + \rho} \tag{20}$$

5.4. Parameter Tuning

The SSA model uses several parameters. As we recall from Section 5.2.2, during the *ssa-mapper* phase, an article is processed to identify salient concepts either based on manual Wikipedia editors' annotations, or on automatically inferred annotations leveraging the one sense per discourse heuristic or the wikification process (methods explained in detail in Section 4.2.1). The wikification process requires two thresholds. The choice of the keyphraseness threshold (probability of hyper-linking a keyword in Wikipedia) was motivated by the size of the corpus. In English, 0.5 was chosen, which is highly restrictive since English is the largest Wikipedia corpus. For this reason a high monosemy threshold of 0.95 (a word is considered to be a keyword only if it refers to a single Wikipedia concept 95% of the time) was further imposed on the candidate keyphrases. In the case of Spanish, Romanian, and Arabic, these conditions were relaxed to correspond with the relative size of the corpus and the nature of the language (Table 4.2.1). Specifically, the threshold for Arabic was lowered to a keyphraseness of 0.2 and a monosemy threshold of 0.65. This

was motivated by the highly inflected nature of the Arabic language when compared to English, Romanian, and Spanish.

5.4.1. Parametric SSA

In order to compute the semantic relatedness score formally defined in Section 4.2.2, based on Equations 17 and 18, the values for the δ, λ, and γ parameters for SSA_p were selected as explained below.

Two tuning datasets are constructed, namely $HM30$ and $HM65$. The datasets are created by replacing $MC30$ and $RG65$ words with synonyms (e.g. replace *lad* with *chap*) or replacing the word-pair with a semantically parallel pair (e.g. replace *bird-crane* with *animal-puma*). Hence, the datasets are similar in regard to the word relations they cover, yet they use completely different terms.

The parameters are adjusted to maximize the correlation on the two tuning datasets. The best matching set of parameters across the two datasets are $\delta = 0.4$, $\lambda = 0.05$, and $\gamma = 0.01$.

It is important to note that the generated tuning datasets do not preserve any word-pairs from the original data, and thus they do not bias the model toward any of the evaluation datasets. This is demonstrated by the moderate correlation figures obtained for the $HM30$ and $HM65$ datasets (see Table 5.4.2) when compared to the high correlation figures for $MC30$ and $RG65$ reported in the literature.

5.4.2. Non-parametric SSA

While using development datasets is useful in estimating the tuning parameters (δ, λ, and γ), it negatively affects the portability and scalability of the model in other languages. Therefore, I also propose a non-parametric version of the SSA model which requires no tuning. This is done by eliminating the γ parameter ($\gamma = 1$). Additionally, the δ value is fixed to 0.3, which allows for maintaining the SSA vectors at a manageable size (Table 5.4.1). For λ normalization, I created a small list of five near synonym word-pairs (Table 5.4.2) along with their translations in Arabic, Romanian, and Spanish. λ was set to the average of the relatedness values of these five pairs, under a given language, which can be calculated at runtime. Since the purpose of λ normalization is to

TABLE 5.4. Frequency vs. Vector Size

Frequency	Vector Size ($\delta = 0.3$)
10	60.38
100	241.53
1000	543.44
10000	966.12
100000	1509.56
1000000	2173.76

close the semantic gap between perfect synonyms (*tiger-tiger*) and near-synonym (*tiger-feline*), we also apply it to the ESA and LSA models (which from our experiments provide better results than their respective un-normalized version).

For the English evaluations, the parametric version of our SSA model will be referred to as SSA_p and the non-parametric version as SSA_n. Mentions of SSA will address both variations. For the multilingual evaluations, I only use the non-parametric version of the model, hence any mentions of SSA will refer to the SSA_n version only.

Table 5.4.2 presents the results obtained using the parametric and non-parametric SSA on the $HM30$ and $HM65$ development datasets. As expected, SSA_n provides lower correlations in comparison to SSA_p, whose parameters were derived based on these datasets.

TABLE 5.5. Related pairs which are used to estimate the value of λ in the non-parametric SSA

English	Romanian	Arabic	Spanish
tiger-lion	tigru-leu	الَاسد-النمر	tigre-león
tiger-feline	tigru-felină	المَاكر-النمر	tigre-felino
fish-whale	peşte-balenă	الحوت-الَاسمَاك	peces-ballena
music-song	muzică-cântec	اغنية-الموسيقَى	música-canción
islam-christian	islam-creştin	مسيحي-الَاسلَام	islam-cristiano

TABLE 5.6. Pearson (r), Spearman (ρ) and their harmonic mean (μ) correlations on the $HM30$ and $HM65$ tuning datasets

	r		ρ		μ	
Metric	**HM30**	**HM65**	**HM30**	**HM65**	**HM30**	**HM65**
SSA_n	0.496	0.639	0.448	0.617	0.471	0.628
SSA_p	0.573	0.670	0.536	0.713	0.554	0.691

CHAPTER 6

EVALUATIONS AND DISCUSSIONS

6.1. English Evaluations

Since the majority of semantic relatedness research in the literature has focused on the English language, various approaches to tackle this problem have been proposed using the selected datasets. For a robust comparison, I picked the top performers to serve as baselines for each of our evaluations. In order to be consistent, the English evaluations were separated into word-relatedness and text-relatedness sections in a fashion similar to the one followed in Section 5.1.1.

6.1.1. Word Relatedness

Table 6.1 shows the results obtained using our tuned (SSA_p) and untuned (SSA_n) salient semantic analysis relatedness model, compared to several state-of-the-art systems: knowledge-based methods including Roget and WordNet Edges (WNE) [35], $H\&S$ [28], $J\&C$ [36], $L\&C$ [44], Lin [52], $Resnik$ [74]; and corpus-based measures such as ESA (as published in Gabrilovich & Markovich [19] and as obtained using our own implementation[1]), LSA [41], and $SOCPMI$ [31]. Excluding the LSA, ESA_{own}, SSA_n, and SSA_p evaluations, which were self conducted, the other reported results are derived from the collected raw data from their respective authors. Some raw data was publicly available in previous publications [34, 35, 50], otherwise it was obtained directly from the authors.

The first examination of the results shows that the knowledge-based methods achieve a high performance for the $MC30$ and $RG65$ datasets, which is probably explained by the deliberate inclusion of familiar and frequently used dictionary words in these sets. The performance quickly degrades on the $WS353$ dataset, largely due to the low coverage: the $WS353$ dataset includes

[1]Since the published ESA results are limited to $MC30$, $WS353$, and $LEE50$, I resolved to use my own ESA implementation to cover the rest of the datasets for a more meaningful comparison. It is worth noting that this latter implementation provides a better Pearson score (0.645) for $MC30$ than the one reported by Gabrilovich & Markovich [19] (0.588) while managing to provide equivalent Pearson scores for $WS353$. Additionally, it outperforms other ESA implementations [91].

TABLE 6.1. Pearson (r), Spearman (ρ) and their harmonic mean (μ) correlations on the English word relatedness datasets

Models	r			ρ			μ		
	MC30	**RG65**	**WS353**	**MC30**	**RG65**	**WS353**	**MC30**	**RG65**	**WS353**
Roget	0.878	0.818	0.413	**0.856**	0.811	0.415	**0.867**	0.814	0.414
WNE	0.732	0.787	0.164	0.768	0.805	0.276	0.749	0.796	0.206
H&S	0.689	0.732	0.335	0.811	0.816	0.348	0.745	0.772	0.342
J&C	0.695	0.731	0.344	0.82	0.806	0.291	0.753	0.767	0.315
L&C	0.821	0.852	0.312	0.768	0.805	0.278	0.793	0.828	0.294
Lin	0.823	0.834	0.341	0.75	0.788	0.348	0.785	0.81	0.344
Resnik	0.775	0.8	0.35	0.693	0.743	0.353	0.732	0.77	0.351
ESA_{Gab}	0.588	NA	0.503	0.727	NA	**0.748**	0.65	NA	**0.602**
ESA_{own}	0.645	0.644	0.487	0.742	0.768	0.525	0.69	0.701	0.506
LSA	0.509	0.45	0.435	0.525	0.499	0.436	0.517	0.473	0.436
SOC_{PMI}	0.764	0.729	NA	0.78	0.741	NA	0.772	0.735	NA
SSA_n	0.771	0.824	0.543	0.688	0.772	0.553	0.727	0.797	0.548
SSA_p	**0.879**	**0.860**	**0.590**	0.843	**0.830**	0.604	0.861	**0.845**	0.597

TABLE 6.2. Pearson (r), Spearman (ρ) and their harmonic mean (μ) correlations on the English text relatedness datasets

Models	r			ρ			μ		
	Li30	**Lee50**	**AG400**	**Li30**	**Lee50**	**AG400**	**Li30**	**Lee50**	**AG400**
ESA_{own}	0.792	0.756	0.434	0.797	0.48	0.392	0.795	0.587	0.412
LSA	0.829	**0.776**	0.4	0.824	**0.523**	0.359	0.826	**0.625**	0.379
Li	0.807	NA	NA	0.801	NA	NA	0.804	NA	NA
STS	0.848	NA	NA	0.832	NA	NA	0.84	NA	NA
SSA_n	0.84	0.744	0.52	0.843	0.371	0.501	0.841	0.495	0.51
SSA_p	**0.852**	0.769	**0.575**	**0.863**	0.434	**0.528**	**0.858**	0.554	**0.551**

proper nouns, technical and culturally biased terms, which are not covered by a typical lexical resource. This factor gives an advantage to corpus-based measures like SSA, LSA, and ESA, which attain the best Spearman results on the $WS353$ dataset.

SSA_p consistently provides superior scores, achieving a 17.5% error reduction for Pearson (r) on the largest dataset ($WS353$[2]) with respect to its closest competitor(ESA_{Gab}). Additionally, it comes second only to ESA_{Gab} Spearman (ρ) score which is the highest reported on the $WS353$ dataset. The harmonic mean of Pearson and Spearman (μ) summarizes the performance of SSA_p and ranks it as the best or the second best across all the datasets, surpassing even the knowledge-based methods. Were we to exclude SSA_p from our discussion, we see that the untuned version (SSA_n) also offers a solid performance. Using Pearson's score, it ranks the third on $RG65$ among knowledge-based methods, and the best on $WS353$ dataset with an error reduction of 8% with respect to its closest competitor (ESA_{Gab}). Generally, SSA_n outperforms all corpus-based and many of knowledge-based models, overcoming even parametric models such as SOC_{PMI}.

6.1.2. Text Relatedness

Table 6.1 presents the text relatedness results for the $Li30$, $Lee50$, and $AG400$ datasets. The results are compared with several state-of-the-art systems: ESA [19], LSA [41], and STS [32]. As seen in Table 6.1, SSA_p and SSA_n obtain the highest correlations for the $LI30$ and $AG400$ datasets[3], even when compared to the STS system, which relies on the SOCPMI framework. While LSA provides the best Pearson score for $Lee50$ ($r = 0.776$), SSA_p comes a close second ($r = 0.769$). It is also interesting to see the large improvements achieved by SSA_p ($\mu = 0.575$) and SSA_n ($\mu = 0.52$) over LSA ($\mu = 0.40$) and ESA ($\mu = 0.434$) when evaluated on the $AG400$ dataset. To explore this in more detail, and also for a comparison with other knowledge-based and corpus-based measures, Table 6.1.2 shows a comparison of the SSA_p and SSA_n with all other relatedness measures reported by Mohler & Mihalcea [64]. As it was the case in the word relatedness evaluations, SSA_p and SSA_n display a performance that is superior to all the

[2]I measured the statistical significance with respect to the largest dataset $WS353$ and concluded that the SSA_p superiority is statistically significant at $p < 0.001$.
[3]SSA's superiority is statistically significant at $p < 0.0005$.

knowledge-based and corpus-based metrics, with an error-reduction of $9.7\% - 17.3\%$ in harmonic mean with respect to the closest competitor ($J\&C$).

Since real-life applications such as Short Answer Grading propose a perfect example of discourse-comprehension, a true semantic model should be able to simulate the comprehension process performed by a grader in understanding and evaluating the discourse. Accordingly, the superior performance demonstrated by the SSA models in this task provides clues about the models' ability to formulate a better world knowledge representation and abstraction that rivals any other knowledge-based or corpus-based representations.

For an additional evaluation of the SSA model, I also measure its ability to recognize paraphrases. To transform SSA into a binary paraphrase classifier, we need to find the similarity threshold k at which the similarity between two candidate sentences is sufficient to classify them as paraphrases of each other. In order to find the best threshold, I use the training set of the MSR dataset and test its accuracy under different thresholds ranging from 0 to 1 in a similar fashion to the STS system reported in Islam & Inkpen [32]. This process is repeated for ESA and LSA.

The systems are evaluated using the traditional Precision (P), Recall (R), F-measure (F), and Accuracy (A) metrics.

$$P = \frac{TP}{TP + FP} \tag{21}$$

$$R = \frac{TP}{TP + FN} \tag{22}$$

$$F = \frac{2PR}{P + R} \tag{23}$$

$$A = \frac{TP}{TP + FP + FN + TN} \tag{24}$$

TABLE 6.3. Comparative results using Pearson (r), Spearman (ρ) and their harmonic mean (μ) for the AG400 dataset, for the relatedness metrics reported in Mohler & Mihalcea [64]

Models	r	ρ	μ
Knowledge-based measures			
WNE	0.440	0.408	0.424
$L\&C$	0.360	0.152	0.214
$Lesk$	0.382	0.346	0.363
$Wu\&Palmer$	0.456	0.354	0.399
$Resnik$	0.216	0.156	0.181
Lin	0.402	0.374	0.388
$J\&C$	0.480	0.436	0.457
$H\&S$	0.243	0.192	0.214
Corpus-based measures			
LSA	0.400	0.359	0.379
ESA_{own}	0.434	0.392	0.412
SSA_n	0.52	0.501	0.51
SSA_p	**0.575**	**0.528**	**0.551**
Baseline			
$tf * idf$	0.369	0.386	0.377

where TP, FP, FN, and TN represent the number of true positives, false positives, false negatives, and true negatives, respectively.

Table 6.4 shows the different accuracies (A) achieved in the training and testing phase of the MSR dataset for STS, SSA_p, SSA_n, LSA, and ESA. All the systems display a consistent performance which peaks between $k = 0.5$ and $k = 0.7$ in both the training and testing phases.

TABLE 6.4. k vs. Accuracy on the MSR dataset

k	STS	SSA_n	SSA_p	ESA	LSA
Training Set (4076)					
0.1	67.54	67.54	67.54	67.54	67.57
0.2	67.54	67.54	67.54	67.59	67.57
0.3	67.59	67.57	67.54	67.62	67.66
0.4	67.74	68.11	67.71	67.52	67.62
0.45	NA	69.46	68.06	67.49	68.06
0.5	69.53	70.04	69.14	**67.76**	68.74
0.55	NA	**71.1**	70.07	67.54	68.96
0.6	**72.42**	70.49	71.57	67.10	**69.26**
0.65	NA	68.89	**71.86**	66.49	69.06
0.7	68.45	66.66	69.23	65.43	68.25
0.75	NA	61.33	65.43	64.50	67.20
0.8	56.67	54.34	58.44	61.33	63.89
0.9	37.78	38.17	40.04	50.98	49.80
1	32.82	33.10	33.19	33.05	34.49
Test Set (1725)					
0.1	66.49	66.49	66.49	66.49	66.55
0.2	66.49	66.49	66.49	66.43	66.55
0.3	66.49	66.49	66.49	66.26	66.78
0.4	66.66	67.48	66.49	66.84	67.25
0.45	NA	70.03	67.19	67.36	67.36
0.5	68.86	71.13	68.00	**67.01**	68.41
0.55	NA	**72.46**	70.32	66.67	**68.17**
0.6	**72.64**	69.80	71.83	65.68	68.81
0.65	NA	67.25	**70.26**	64.99	69.16
0.7	68.06	64.87	68.46	64.70	68.29
0.75	NA	61.22	64.17	63.19	67.13
0.8	56.29	54.03	57.86	60.81	65.51
0.9	38.38	40.00	41.10	50.49	51.48
1	33.79	34.03	34.09	33.91	35.36

The first observation is that STS performs slightly better than SSA_n and SSA_p. This is expected when taking into consideration that the MSR dataset was constructed using the Levenshtein distance ($n \leq 12$), along with some journalistic heuristics [16] which removed sentences that did not share at least three words longer than four characters. These heuristics grant a greater advantage to STS, which incorporates the Levenshtein distance.

To get a wider perspective regarding SSA's performance in paraphrase detection, I compare it to the state-of-the-art systems reported in [1, 32, 58] (see Table 6.5). The lexical-based category in the table refers to systems that adapt lexical matching techniques, be they through simple overlap $\{Lex\}$, normalized overlap $\{Jaccard\}$, weighted overlap $\{Lex_{IDF}, Lex_{identity}, Lex_{novelty}\}$, phrasal overlap $\{Lex_{phrase}\}$, order sensitive overlap $\{Li_{order}\}$, or just cosine distance $\{Lex_{cosine}\}$. In addition, hybrid systems that mix knowledge-based and corpus-based approaches $\{Li_{SV}, Li_{SV+Order}\}$ are also included. For an overview of these systems the reader is advised to consult Achananuparp et al. [1]. The SSA accuracies ($70.3\% - 72.5\%$) are among the highest achieved across all lexical and knowledge-based systems and come only second to STS (72.6%). Also, the F-measure obtained by SSA_n (81.4%) ranks it as the best performer. Interestingly, the untuned version of SSA_n outperforms the tuned version SSA_p. This might be due to tuning bias toward relatedness tasks versus paraphrase detection imposed by the development datasets.

Similar to the Short Answer Grading task, paraphrase detection requires inference and knowledge about the world. Hence it requires that the reader achieves a situational model level of comprehension in order to recognize and infer these parallel meanings. The SSA performance in this task reaffirms its ability to simulate our abstraction and representation of world knowledge in an effective fashion.

6.2. Multilingual Evaluations

In this section we explore the hypothesis that incorporating information/knowledge from multiple monolingual models may lead to a stronger semantic relatedness. I also examine whether the performance of the salient semantic analysis models holds under different monolingual and multilingual settings. Consequently, I am seeking to answer the following questions:

TABLE 6.5. MSR Results

Models	A	P	R	F
corpus-based				
$PMI\text{-}IR$	69.9	70.2	95.2	81.0
$STS(0.6)$	**72.6**	74.7	89.1	<u>81.3</u>
$SSA_u(0.55)$	<u>72.5</u>	73.9	90.7	**81.4**
$SSA_t(0.65)$	70.3	75.2	82.4	78.7
$ESA(0.5)$	67.0	68.0	95.2	79.3
$LSA(0.6)$	68.8	70.0	92.9	79.9
knowledge-based				
$J\&C$	69.3	72.2	87.1	79.0
$L\&C$	69.5	72.4	87.0	79.0
$Lesk$	69.3	72.4	86.6	78.9
Lin	69.3	71.6	88.7	79.2
$W\&P$	69.0	70.2	92.1	80.0
$Resnik$	69.0	69.0	<u>96.4</u>	80.4
hybrid-models				
Li_{SV}[50]	66.8	66.9	98.9	79.8
$Li_{SV+Order}$[50]	67.1	67.3	98.3	79.9
lexical-models				
$Jaccard$	65.7	<u>83.5</u>	60.3	70.0
Lex	64.3	76.0	67.8	71.7
Lex_{IDF}[57]	50.7	82.9	32.5	46.7
Lex_{phrase}[70]	67.5	70.0	89.2	78.5
$Lex_{novelity}$[2]	49.2	**85.8**	28.3	42.6
$Lex_{identity}$[29]	66.4	66.5	**100.0**	79.8
Li_{order}[50]	55.4	68.1	61.9	64.8
Lex_{cosine}	65.4	71.6	79.5	75.3
$Random$	51.3	68.3	50.0	57.8

- Does incorporating additional information/knowledge from different languages allow for a better semantic relatedness abstraction?
- Do some monolingual models benefit from using this abstraction more than others?
- Does this multilingual assumption hold under automated settings?
- Does SSA maintain its superiority under the multilingual model paradigm?

Does incorporating additional information/knowledge from different languages allow for a better semantic relatedness abstraction?

To answer this question, I create SSA, ESA and LSA systems for Romanian, Arabic, and Spanish, using the same Wikipedia versions for each of the systems. These models are evaluated on the manually constructed multilingual datasets ($MC30$, $WS353$, and $LI30$) described in Section 5.1.2.

To construct a multilingual model, I start with a word/text-pair from a source language along with its translations in the other languages. Then, the relatedness scores achieved for this pair under the different monolingual models are simply aggregated to form a final relatedness score, where the adopted aggregation function utilizes a simple average.

To evaluate this multilingual model in a fashion that would reduce the bias that may arise from choosing one language over the other, the following steps are undertaken: first, starting from a source language, all the possible combinations of this language with the available language set $\{ar, en, es, ro\}$ are generated. Then, within each combination, the monolingual model scores for the languages in this combination with respect to the target word/text pair are aggregated into a final relatedness score.

For example, let us consider Spanish as the source language, then the possible combinations of the languages that include the source language will be $\{\{es\}$, $\{es, ar\}$, $\{es, ro\}$, $\{es, en\}$, $\{es, ar, en\}$, $\{es, ar, ro\}$, $\{es, en, ro\}$, and $\{es, ar, en, ro\}\}$. For each possible combination, the scores of the languages in that combination are aggregated. In this setting, a combination of size (cardinality) one will always be the source language and will serve as the baseline. For every combination (e.g. $\{es, ar\}$), the individual monolingual relatedness scores for a given word/text-pair in this set are averaged.

Finally, to calculate the overall correlation of these generated multilingual models (one system per combination size) with the human scores, I average the correlation scores (r, ρ, μ) achieved over all the datasets in a given combination (e.g. $\{es, ar\}$) with all correlation scores achieved under other combinations of the same size (e.g. $\{es, ro\}$, $\{es, en\}$). This in effect allows us to observe the cumulative performance irrespective of language choice, as the multilingual model is extended to include more languages.

Formally, let N be the number of languages, C_n be the set of all language combinations of size n, and c_i be one of the possible combinations of size n,

$$C_n = \{c_i \mid |c_i| = n, 0 < i < \binom{N}{n}\} \tag{25}$$

then the relatedness of a word/text pair p from the dataset P under this combination can be represented as:

$$Sim_{c_i}(p) = \frac{1}{|c_i|} \sum_{l \in c_i} Sim_l(p) \tag{26}$$

where $Sim_l(p)$ is the relatedness score of the word/text pair p in the monolingual model of language l. To evaluate the performance of the multilingual model, let D_i be the generated relatedness distribution for the dataset P using the combination c_i:

$$D_i = \{\langle p, Sim_{c_i}(p)\rangle \mid p \in P\}. \tag{27}$$

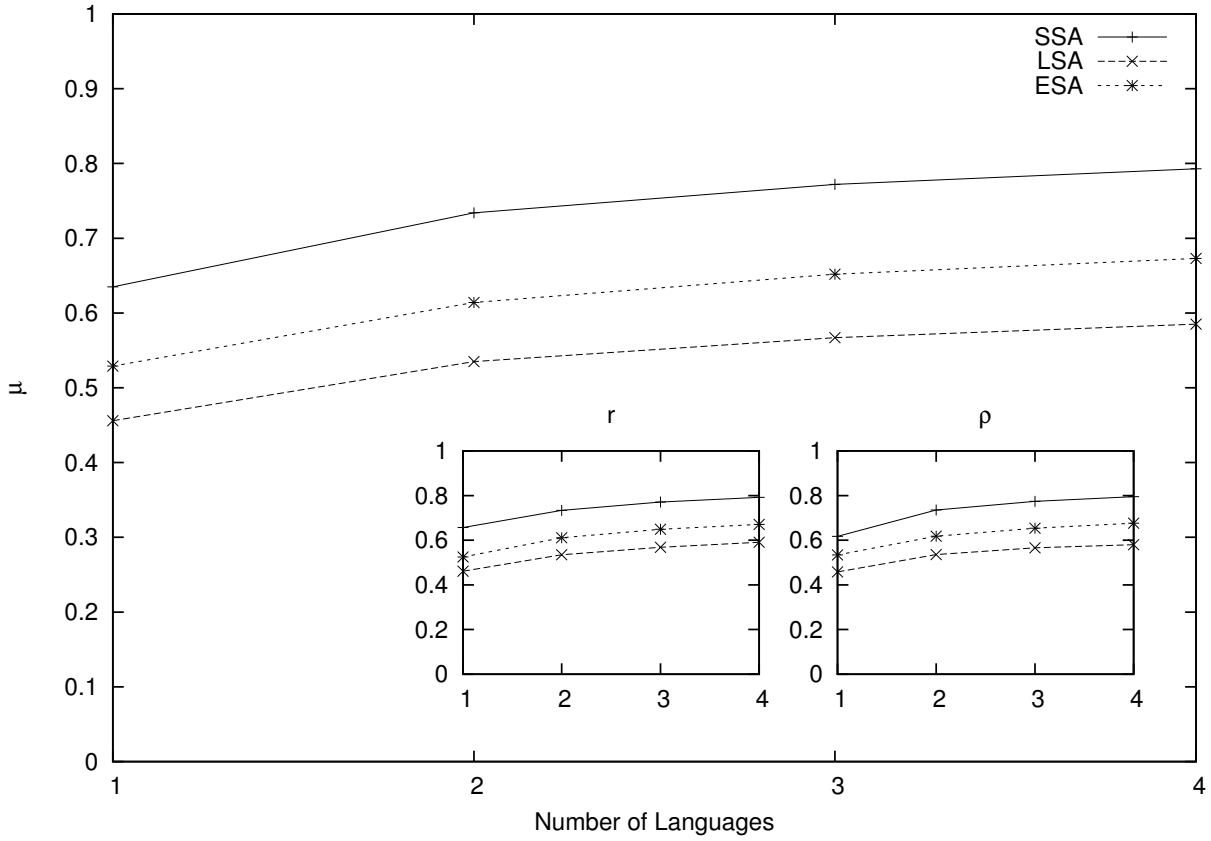
Then, the correlation between the gold standard distribution G and the generated scores can be calculated as follows:

$$Correlation_{C_n}(D, G) = \frac{1}{|C_n|} \sum_{c_i \in C_n} Correlation_{c_i}(D_i, G), \tag{28}$$

where $Correlation$ can stand for Pearson (r), Spearman (ρ), or their harmonic mean (μ).

By plotting the correlation scores achieved across all the languages and then averaged across all the multilingual datasets in Figure 6.1, a clear and steady improvement (25% - 28% with respect to the monolingual baseline) is achieved when incorporating more languages. It is

FIGURE 6.1. Using manual translations, how systems' performance on average benefits from incorporating scores from models in other languages

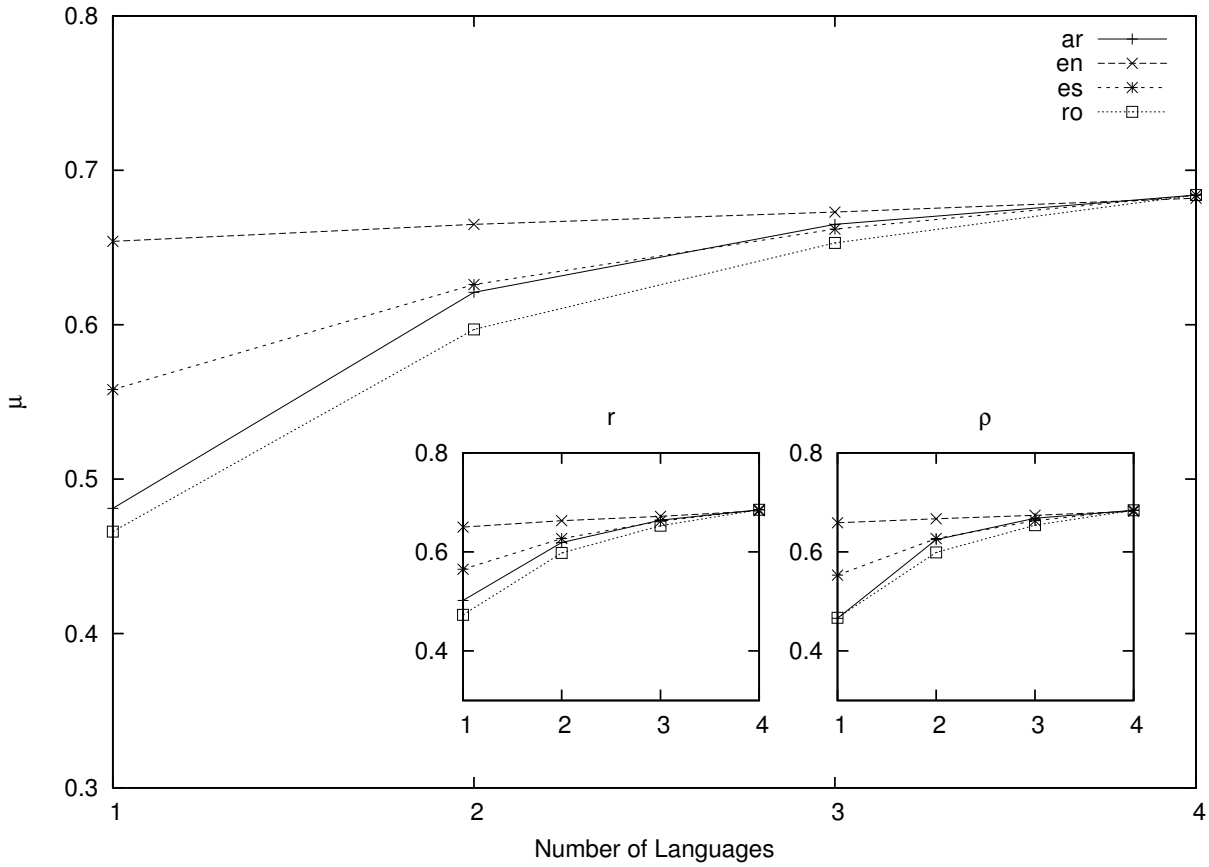


worth noting that both Pearson and Spearman correlations exhibit the same improvement pattern which confirms our hypothesis that adding more languages has a positive impact on the achieved point-wise relatedness scores, as well as on the overall ranking of these scores as observed from their harmonic mean (μ). Additionally, SSA maintains its performance and consistency when compared with ESA and LSA.

Do some monolingual models benefit from using this abstraction more than others?

To further analyze the role of the multilingual model and to explore whether some languages benefit from using this abstraction more than others, the correlation scores achieved by the individual languages averaged over all the systems and the datasets are plotted in Figure 6.2. A sharp rise in performance associated with the addition of more languages to the Arabic (42%)

FIGURE 6.2. Using manual translations, how models in source languages benefit from incorporating information from other languages



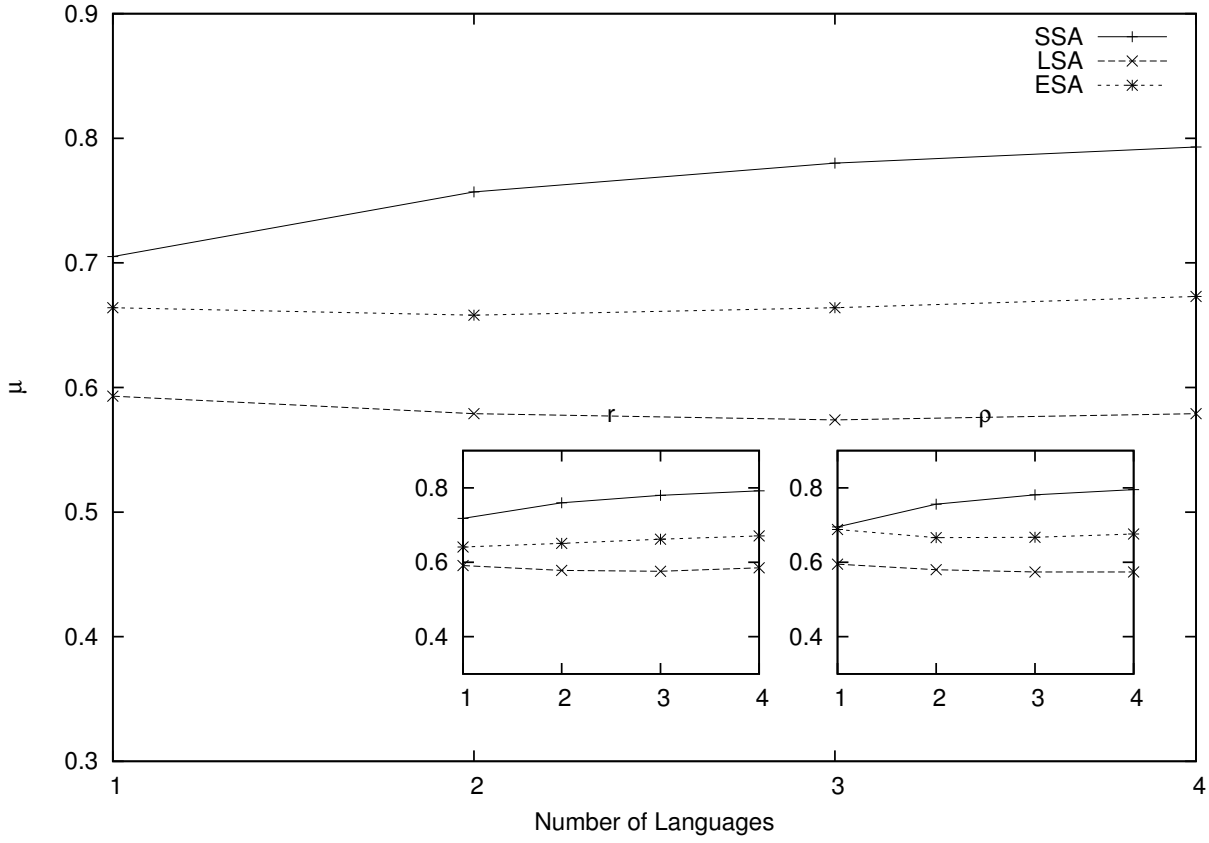
and the Romanian (47%) models is noticed, while Spanish exhibits a slower rise (23%). The performance of English is also affected, but on a smaller scale (4%) when compared to the other languages. Interestingly, this inversely correlates ($Pearson = -0.85$) with the size of each corpus (Table 4.2.1), where Arabic and Romanian are the smallest, while English is the largest. Moreover, the inverse correlation is even stronger ($Pearson = -0.93$) with the number of manually disambiguated concepts extracted from Wikipedia (Table 4.2.1), which reflects the amount of information leveraged by SSA.

Motivated by this observation, I perform a variation of this evaluation using a weighted average, where the monolingual relatedness scores are weighted by the log of the number of manually disambiguated concepts discovered in the corresponding monolingual corpus. This evaluation

however leads to only a small improvement. This might be caused by the fact that the sizes of the Romanian and Arabic models are already reflected in the smaller point-wise relatedness they produce in comparison to English and Spanish.

Not surprisingly, the observed pattern matches perfectly with the reported pattern in Banea & Mihalcea [4] when progressively evaluating the addition of features from different languages to generate a supervised system for subjectivity analysis. The results support the notion that projecting models from a resource poor language into a language with richer and larger resources, such as English or Spanish, leads to a better performance. Furthermore, incorporating additional languages to English also leads to small improvements, which indicates that the benefit, while disproportionate, is mutual.

FIGURE 6.3. Using manual translations, how English models performance on average benefits from incorporating scores from models in other languages



To investigate this even further, English is analyzed into a separate plot (Figure 6.3) in which the evaluation is broken down by the individual systems. Looking at the individual performance of each system as we add more languages to English, we see that LSA suffers a small loss of 2%, however it is overshadowed by the 12.4% improvement in SSA. ESA, on the other hand, experiences a 1.5% improvement.

Does this multilingual assumption hold under automated settings?

To take the previous discussion to its natural conclusion, I further investigate the role of the manual translations in the multilingual model performance. Since the previous evaluations require the availability of the word/text pairs in multiple languages - a hard requirement to satisfy -, I attempt to see if this restriction can be eliminated by automating the translation process using statistical machine translation (MT). Therefore, for a multilingual model employing automated settings, the manual models proposed previously constitute an upper bound.

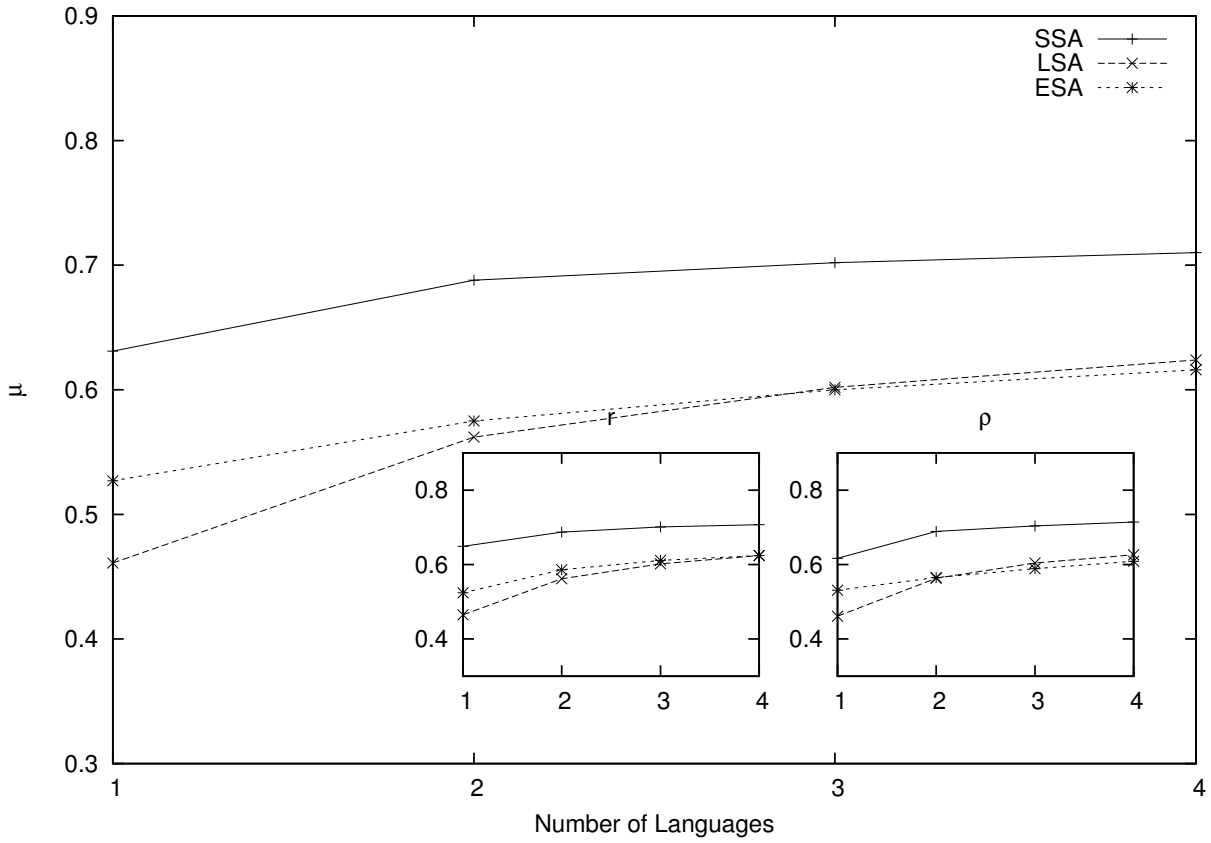
Thus, I use the Google MT engine[4] to translate the multilingual datasets into the target languages (en, es, ar, and ro). All the evaluations are then repeated using the newly constructed datasets.

Figure 6.4 represents the correlation scores achieved across all the languages and averaged across all the multilingual datasets constructed using automatic translation. There is again a clear and steady improvement (12% - 35% with respect to the monolingual baseline) similar to the observed pattern in the corresponding manual evaluations (Figure 6.1). While the overall achieved performance for SSA has dropped (from $\mu = 0.793$ to $\mu = 0.71$) when compared to the manual settings, a large improvement over the baseline (from $\mu = 0.635$) is still achieved[5]. LSA seems to experience the highest relative improvement (35%), which might be due to its ability to handle noise in these automatic settings. Overall Pearson and Spearman correlations exhibit the same improvement pattern, which supports the notion that even with the possibility of introducing noise through mistranslations, the models overall benefit from the additional clues.

[4] `http://translate.google.com/`

[5] Interestingly, a relatively small drop in performance is also reported by Banea et al. [5] when switching from manual to automatic translations for subjectivity classification.

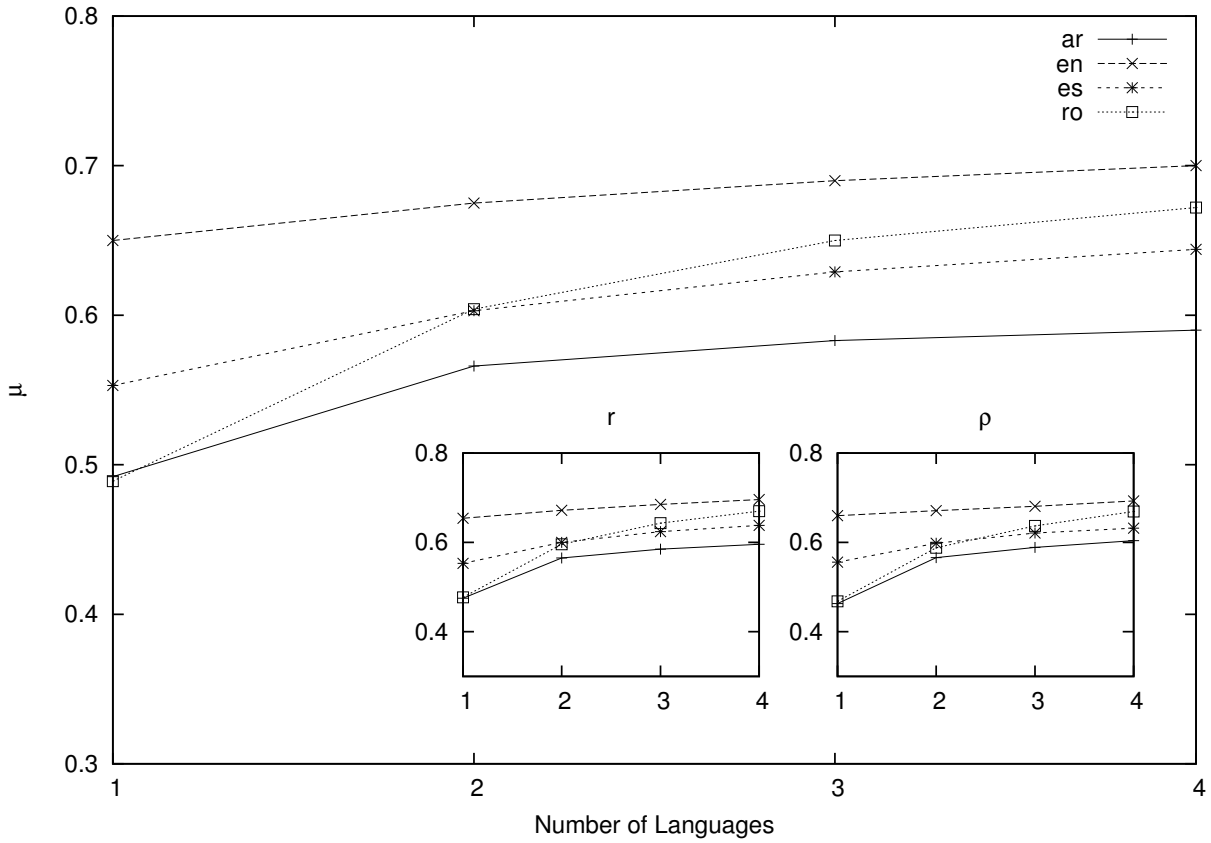
FIGURE 6.4. Using automatic translations, how systems' performance on average benefits from incorporating scores from models in other languages



To explore the effect of automatic translations on the individual languages, we plot the correlation scores achieved vis-à-vis a reference language, and average over all the systems and the automatically translated datasets in Figure 6.5, in a similar fashion to Figure 6.2.

We notice the similar rise in performance associated with the addition of more languages to the Arabic (20%) and the Romanian (37%) models, and a slower rise for Spanish (16%) and English (8%). The effect of the automatic translation quality is evident for the Arabic language where the automatic translation seems to slow down the improvement when compared to manual translations (Figure 6.5). A similar behavior is also observed in Spanish and Romanian, but on a lower scale.

FIGURE 6.5. Using automatic translations, how models in source languages benefit from incorporating information from other languages

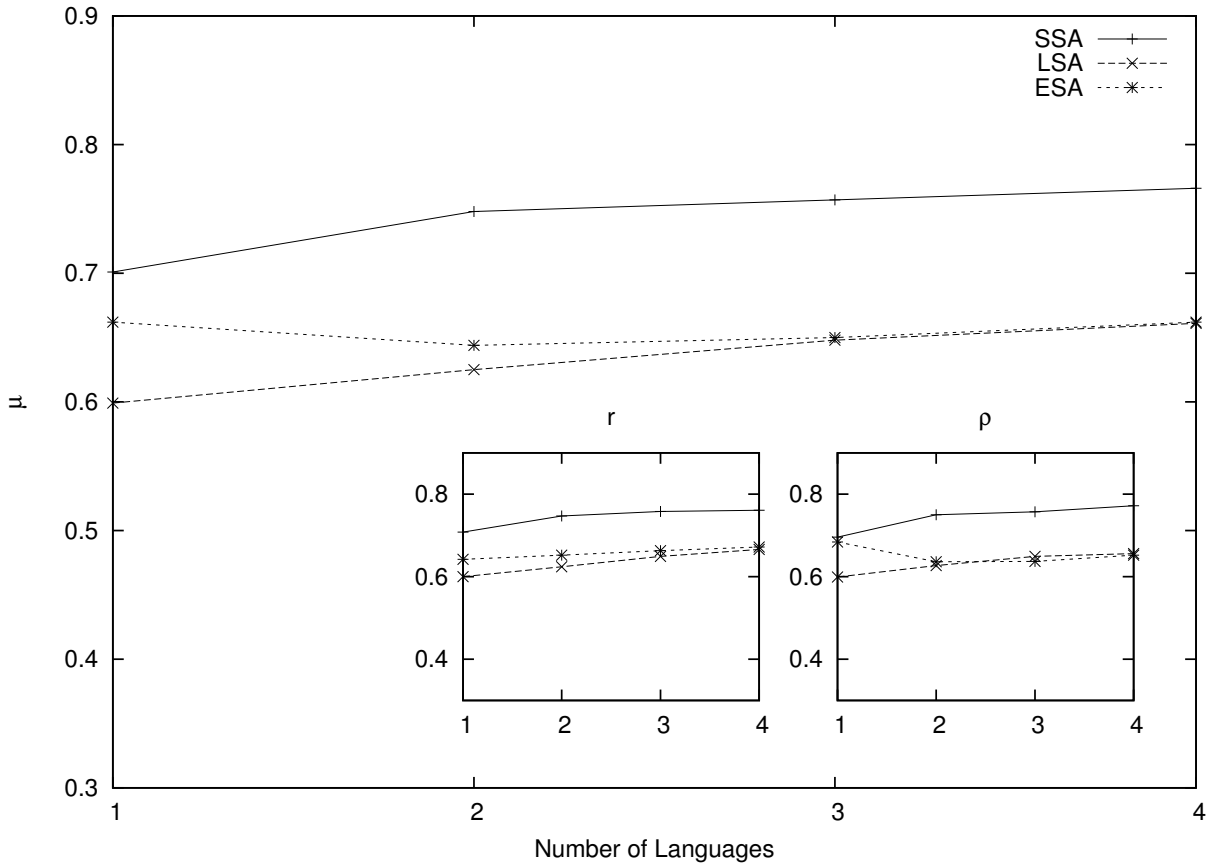


A very interesting consideration is that English experiences a stronger improvement when using automatic translations (8%) compared to manual translations (4%). This can be attributed to the translation engine quality in transferring English text to other languages and to the fact that the statistical translation (when accurate) can lead to a translation that makes use of more frequently used words, which contribute to more robust relatedness measures. When presented with a word pair, human judges may provide a translation influenced by the form/root of the word in the source language, which may not be as commonly used as the output of a MT system. For example, when presented with the pair "coast - shore," a Romanian translator may be tempted to provide "coastă" as a translation candidate for the first word in the pair, as it resembles in form the English word. However, the Romanian word is highly ambiguous, and in an authoritative

Romanian dictionary[6] its primary sense is that of rib, followed by side, slope, and ultimately coast. Thus, a MT system using a statistical inference may provide a stronger translation such as "ţărm" that is far less ambiguous, and whose primary meaning is the one intended by the original pair.

Overall, the improvement trend is positive and follows the trends previously observed on the manually constructed datasets. This suggests that an automatic translation, even if more noisy, is beneficial and provides a unique opportunity to reinforce semantic relatedness in a given language with information coming from multiple languages with no manual effort.

FIGURE 6.6. Using automatic translations, how the average performance for the English model benefits from incorporating scores from models in other languages



[6] http://dexonline.ro/definitie/coasta

Additionally, I investigate the English based multilingual models further by considering them separately in Figure 6.6. This evaluation, much like the one presented in Figure 6.3, represents the performance of the different systems with respect to the additional languages. We notice that ESA's performance suffered a loss of less than 1%. This seems to be in large part due to the $Spearman$ performance. SSA and LSA, on the other hand, scored an improvement of 10%. This could be explained by the fact that context expansion might have occurred in the automatic translation (e.g. some words are translated into multiple words), which was not allowed in the manual generation of translations for word-to-word datasets. Overall, the loss due to the use of MT seems to be reasonable and does not affect the integrity of the multilingual model.

At last, to see how this multilingual model can improve over the evaluations performed in the earlier English experiments, $LEE50$, $RG65$, and $AG400$ datasets were automatically translated into all the languages in our language set and all the Word-to-Word and Text-to-Text evaluations were repeated, as seen in the following sections.

Does SSA maintain its superiority under the multilingual model paradigm?

By analyzing Figures 6.1, 6.3, and 6.6, we remark that SSA is consistent in its performance across all the monolingual baselines ($Number\ of\ languages$ = 1), and its performance increases steadily, for every additional language participating in a combination. This, I believe, is due to the ability of SSA to leverage much more information from Wikipedia then ESA and LSA, which results in a richer representation.

6.2.1. Word Relatedness

The word relatedness experiments performed in Section 6.1.1 are repeated by incorporating the multilingual model that includes all languages.

Table 6.2.1 shows the correlation scores achieved with automatic translation. The positive influence of the multilingual model is emphasized by the clustering of the top scores in the multilingual version of each system. Specifically, the SSA model experiences an improvement in μ of 26% for $WS353$ and 15% for $MS30$. This improvement is most evident in the case of the largest dataset $WS353$, where all the multilingual models exhibit a consistent and strong performance.

TABLE 6.6. Pearson (r), Spearman (ρ) and their harmonic mean (μ) correlations on the word relatedness datasets using multilingual models

Models	r			ρ			μ		
	MC30	**RG65**	**WS353**	**MC30**	**RG65**	**WS353**	**MC30**	**RG65**	**WS353**
ESA_{en}	0.645	0.644	0.487	0.742	**0.768**	**0.525**	0.690	0.701	0.506
ESA_{ml}	**0.723**	**0.741**	**0.515**	**0.766**	0.759	0.519	**0.744**	**0.75**	**0.517**
LSA_{en}	0.509	0.450	0.435	**0.525**	0.499	0.436	**0.517**	0.473	0.436
LSA_{ml}	**0.538**	**0.566**	**0.487**	0.484	**0.569**	**0.517**	0.510	**0.567**	**0.502**
SSA_{en}	0.771	**0.824**	0.543	0.688	0.772	0.553	0.727	0.797	0.548
SSA_{ml}	**0.873**	0.807	**0.674**	**0.803**	**0.795**	**0.713**	**0.836**	**0.801**	**0.693**

6.2.2. Text Relatedness

In the Text-to-Text evaluations, the multilingual model demonstrates a similar performance to the Word-to-Word evaluations (see Table 6.2.2). While the ESA performance suffers in the multilingual model, it is overshadowed by the improvement experienced by LSA and SSA. The multilingual model reports some of the best scores in the literature, such as a correlations of $r = 0.856$ and $\rho = 0.87$ for $LI30$ achieved by LSA and SSA, respectively. Not surprisingly, SSA is still a top contender, achieving the highest scores for $AG400$ and $LI30$. In $AG400$, SSA reports a μ of 0.53 which represents a 4% improvement over the English SSA model ($\mu = 0.51$) and a 16% improvement over the best knowledge-based system $J\&C$ ($\mu = 0.457$).

TABLE 6.7. Pearson (r), Spearman (ρ) and their harmonic mean (μ) correlations on the text relatedness datasets using multilingual models

Models	r			ρ			μ		
	LI30	**LEE50**	**AG400**	**LI30**	**LEE50**	**AG400**	**LI30**	**LEE50**	**AG400**
ESA_{en}	**0.792**	**0.756**	**0.434**	**0.797**	**0.48**	**0.392**	**0.795**	**0.587**	**0.412**
ESA_{ml}	0.776	0.648	0.382	0.742	0.339	0.358	0.759	0.445	0.369
LSA_{en}	0.829	**0.776**	0.400	0.824	**0.523**	0.359	0.826	**0.625**	0.379
LSA_{ml}	**0.856**	0.765	**0.46**	**0.855**	0.502	**0.404**	**0.856**	0.606	**0.43**
SSA_{en}	**0.840**	**0.744**	0.520	0.843	0.371	0.501	0.841	0.495	0.510
SSA_{ml}	0.829	0.743	**0.539**	**0.87**	**0.41**	**0.521**	**0.849**	**0.528**	**0.53**

CHAPTER 7

CONCLUSION

I will now refer back to the contributions proposed in Section 1.3, and how this work addressed them.

7.1. Salient Concepts

Propose a new interpretation of the semantic context using salient concepts identified in a context's lexicon.

Identifying the correct semantic interpretation of context is very valuable in understanding natural language. I investigated the role of a special lexicon tagged as "concepts" in the semantic modeling of context and the way this could be interpolated to generate a semantic interpretation of all lexical units may they be words, sentences, or documents.

The newly proposed salient semantic analysis draws its inspiration from the psycholinguistic notion of concept as an atomic mental unit which serves as the basis for discourse comprehension. In this representation propositions are regarded as mere constraints on concepts and characterize the smallest unit of meaning that asserts information about the world. Furthermore, propositions are truth bearers, in the sense that they resolve to be either true or false.

The model rearticulates this interpretation by defining a subset of unambiguous and well defined concepts as *salient*. These salient concepts are important in understanding and disambiguating the context and have the ability to easily integrate in our working memory, which was supported by the recall experiment we performed (Table 4.2).

By using these salient concepts to index the surrounding context, the system attains a robust and rich concept-based semantic representation that rivals that of state-of-the-art semantic models.

The concept based model proposed here is grounded in psycholinguistic research, and its assumptions align with state-of-the-art psycholinguistic theories (Chapter 4). Additionally, I motivate the reason behind selecting an expository text type such as Wikipedia, as it allows for the inference of a model embedded with a coherent world representation.

7.2. Concept-Based Semantic Representations

Examine the salient concept-space representation of meaning as more effective and robust when compared to contending representations.

Concept-based representations have been widely successful in modeling the semantic interpretation of text. Its theoretical background is grounded in the psycholinguistics domain and supported by various studies. Kintsch [39] argues that such concept-based representation, specifically latent semantic analysis, represents the basis of semantic theory and can account for the associative representation of knowledge.

My evaluations have covered two of the most popular concept-based representations, namely latent semantic analysis and explicit semantic analysis, along with the newly proposed salient semantic analysis model. While these models differ algorithmically, their output is similar in the way it reduces the context into a set of relevant concepts. While LSA leverages latent concepts in raw text based on co-occurrence information, ESA is able to infuse a vertical structure (Wikipedia article information) over what could otherwise be viewed as raw text. SSA disregards Wikipedia article structure and only utilizes annotations in harvesting salient features, thus focusing on a horizontal textual representation. This allows the model to be scalable to unstructured text in the future.

The conducted experiments have demonstrated that the salient semantic analysis model is highly competitive in the semantic relatedness task, as well as in real-life applications such as short answer grading and paraphrase detection, all of them being tasks that require a high level of comprehension, inference, and background world knowledge. This performance surpasses that of all other knowledge-based models as well, due to the unsupervised nature of SSA and its ability to operate in uncontrolled vocabulary settings.

7.3. Language Independent Semantic Relatedness

Explore the association between the semantic relatedness of context and the choice of communicating language.

While most of the previous studies of semantic relatedness confined the scope of the semantic models to a specific language (mostly English), in this work the angle is broadened to multiple languages. Specifically, I explored the portability of semantic relatedness across languages under controlled settings and whether it is affected by the choice of the target language or the subjectivity of the annotator.

To this end, a set of annotation experiments were performed to assess the transfer of the semantic relatedness of textual units from one language to another. Experiments conducted on newly constructed fine-grained word-to-word (Section 5.1.2.1) and coarse-grained text-to-text (Section 5.1.2.2) datasets for Spanish and Arabic demonstrate a high correlation between the relatedness scores assigned by native speakers of different languages on the same dataset (presented in different languages). This entails that the semantic relatedness can overcome the ambiguity inherent in language and successfully cross language boundaries. Furthermore, it follows that the mental representation of meaning is not only consistent but universal.

7.4. Multilingual Semantic Relatedness

Propose a new scheme of incorporating monolingual semantic models in a multilingual setting to improve semantic relatedness.

Once having established that semantic relatedness can be consistent across language boundaries, the natural course was to harvest clues from multiple monolingual models, and seek to aggregate them, thus allowing each language to participate in a reinforcement process, at the end of which a strengthened, clearer semantic relatedness would transpire.

By applying this multilingual generalization to latent semantic analysis, explicit semantic analysis and our salient semantic analysis, a steady improvement is observed over a simple monolingual baseline which strongly correlates with the number of languages involved. Specifically, the proposed method achieved some of the best reported scores in the literature for our evaluation datasets, namely $WS353$, $LI30$, $AG400$, and $LEE50$.

The evaluations showed that some languages benefit from this abstraction more than others. For example, languages like Arabic and Romanian experience a large improvement when adding semantic clues from resource rich languages such as English and Spanish. This behavior strongly correlates with the size of the corpora available for these languages.

Overall, the results support the notion that reinforcing models from a resource poor language with clues originating from a language with richer and larger resources, such as English or Spanish, leads to an improvement in the semantic relatedness performance. This behavior is not limited though to only languages with few electronic resources. Incorporating additional languages to an English monolingual model also leads to incremental improvements in performance, which indicates that the benefit, while disproportionate, is mutual.

This improvement is largely maintained even when incorporating statistical machine translation to automate the generation of multilingual context representations. While manual translations entail a better performance, the semantic relatedness task under these settings is more synthetic and can truly act only as an upperbound; under the automatic translation paradigm the multilingual semantic relatedness model is fully automated and can easily be extrapolated to include information originating from a large number of languages without any human intervention beyond building the monolingual vectorial spaces used by either ESA, LSA or SSA.

Additionally, the salient semantic analysis model displays a superior, consistent, and robust performance across the multilingual evaluations which indicates its ability to maximize the use of the underlying corpora in extracting feature rich representations irrespective of the chosen language.

7.5. Multilingual Evaluation Framework

Propose a framework for evaluating semantic relatedness in multilingual settings.

Since this research requires the utilization of multiple monolingual semantic models covering a diverse set of languages, I had to establish a robust evaluation framework to accommodate these settings. This includes formally defining and standardizing the process of datasets construction (Section 5.1.2), the evaluation metrics (Section 5.3), and the assessment strategy (Section 6.2). I believe that the framework was successful in enforcing equivalent settings so that fully

comparable monolingual models (without bias to the choice of languages) would play a role in the multilingual models' behavior.

7.6. Future Work

Currently, I am exploring the use of the salient semantic analysis model for query expansion in the Information Retrieval framework. The basic idea is to construct a salient semantic context vector from a given query, then utilize the titles of the top ranked salient concepts in the vector to enrich the query. Initial evaluations demonstrate robust and superior performance for SSA when compared to published results [71].

Additionally, since some salient concepts can be synonymous, it would be interesting to explore the use of dimensionality reduction algorithms to condense the salient concept space to a more compact form. For example, considering the "automobile" example (see Table 4.2.2.1), we would like to compress similar/synonymous salient concepts (e.g. the top seven concepts) into a single representative concept.

Additionally, since SSA is not dependent on the underlying structure of Wikipedia, I believe it can be ported to unstructured text. Due to the lack of user annotations in such settings, I believe we can utilize automated approaches [59] to identify and tag these salient concepts, thus allowing for the generalization of SSA beyond Wikipedia boundaries.

Appendices

APPENDIX

Table A.1: HM30 Dataset

Word	Word	$score$
crazy	madhouse	3.26
animal	mouse	2.8
animal	puma	3.1
cadet	chap	3.08
fellow	priest	2.28
vehicle	auto	3.3
graveyard	forest	1.04
thread	grin	0
shore	woodland	1.2
shore	mound	0.8
shore	beach	3.78
bulldozer	equipment	2.52
meal	vegetable	2.64
meal	chicken	2.74
woodland	cemetery	1.4
stove	oven	3.44
treasure	pearl	3.12
cup	wizard	0.4
trip	automobile	2.12

Continued on next page

Table A.1 – continued from previous page

Word	Word	*score*
pilgrimage	trip	3.3
chap	buddy	3.1
chap	magician	1.2
wizard	sorcerer	4
midnight	late	3.24
priest	prophet	2.94
priest	captive	0.4
midnight	thread	0
chicken	journey	0.2
beach	forest	1.1
instrument	device	3.5

Table A.2: HM65 Dataset

Word	Word	*score*
asylum	tomb	0.3
asylum	harvest	0.1
crazy	asylum	3.26
asylum	priest	0.8
handwrite	beach	0
handwrite	inscribe	3.38
vehicle	auto	3.9
vehicle	soft	0.3
vehicle	magician	0.2

Continued on next page

Table A.2 – continued from previous page

Word	Word	*score*
animal	mouse	3.16
animal	puma	3.16
cat	forest	1.375
child	chap	3.14
child	duck	0.5
child	guru	1
chap	buddy	3.3
fellow	priest	2.34
trip	automobile	3.06
tomb	burial	3.5
tomb	hill	2.26
graveyard	timberland	1.08
shore	timberland	0.94
shore	mound	0.64
shore	beach	3.58
puma	cat	3.58
thread	grin	0
thread	strand	3.6
bulldozer	equipment	2.96
puma	tiger	3.38
soft	treasure	0.2
sofa	couch	3.9
meal	vegetable	2.92
meal	rooster	2.72
timberland	cemetery	1.6

Continued on next page

Table A.2 – continued from previous page

Word	Word	*score*
jungle	forest	3.72
grapes	fire	0
stove	instrument	2.5
stove	oven	3.72
treasure	pearl	3.06
crystal	pearl	3.26
cup	wizard	0.84
cup	grail	3.7
tomb	asylum	0.5
smile	instrument	0
smile	cadet	0.7
smirk	happy	3.32
climb	mountain	3.5
mountain	forest	2.6
instrument	device	3.6
pilgrimage	trip	3.46
chap	magician	1.26
witch	prophet	1.56
wizard	sorcerer	3.9
midnight	late	3.04
priest	prophet	3.2
priest	captive	1.06
hill	beach	1.36
dune	furnace	0.04
midnight	thread	0.1

Continued on next page

Table A.2 – continued from previous page

Word	Word	$score$
prophet	teacher	2.84
chicken	journey	0.14
guru	magician	1.5
servant	worker	3.76
coast	journey	2.5
beach	forest	1.06

BIBLIOGRAPHY

[1] Palakorn Achananuparp, Xiaohua Hu, and Xiajion Shen, *The evaluation of sentence similarity measures*, Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery (DaWaK'08) (Turin, Italy), 2008, pp. 305–316.

[2] James Allan, Courtney Wade, and Alvaro Bolivar, *Retrieval and novelty detection at the sentence level*, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03 (2003), 314.

[3] John R. Anderson, *Concepts, propositions, and schemata: what are the cognitive units?*, Nebraska Symposium on Motivation, Volume 28: Human Cognition (Lincoln, Nebraska, US.), vol. 28: Human, University of Nebraska Press, 1980, pp. 121–162.

[4] Carmen Banea and Rada Mihalcea, *Word sense disambiguation with multilingual features*, Proceedings of the 9th International Conference on Computational Semantics (IWCS'11) (Oxford, UK), 2011, pp. 25–34.

[5] Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan, *Multilingual subjectivity analysis using machine translation*, Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP'08) (Honolulu, Hawaii), 2008, pp. 127–135.

[6] Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena, *International sentiment analysis for news and blogs*, Proceedings of the International Conference on Weblogs and Social Media (ICWSM'08) (Seattle, Washington), 2008.

[7] Romaric Besançon and Martin Rajman, *Evaluation of a vector space similarity measure in a multilingual framework*, Proceedings of the 3rd International Conference on Language Resource and Evaluation (LREC'02) (Las Palmas, Spain), 2002.

[8] Sergey Brin, James Davis, and Héctor García-Molina, *Copy detection mechanisms for digital documents*, Proceedings of the 1995 ACM SIGMOD international conference on management of data (ACM SIGMOD'95) (San Jose, California, United States), vol. 24, May 1995, pp. 398–409.

[9] Andrei Z. Broder, *Syntactic clustering of the Web*, Computer Networks and ISDN Systems 29 (1997), no. 8-13, 1157–1166.

[10] Andrei Z. Broder, Peter Ciccolo, Marcus Fontoura, Evgeniy Gabrilovich, Vanja Josifovski, and Lance Riedel, *Search advertising using web relevance feedback*, Proceeding of the 17th ACM conference on Information and knowledge mining (CIKM '08) (New York, New York, USA), ACM Press, 2008, pp. 1013–1022.

[11] Alexander Budanitsky and Graeme Hirst, *Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures*, Workshop on WordNet and Other Lexical Resources, Second meetings of the North American Chapter of the Association for Computational Linguistics (Pittsburgh, Pennsylvania, US.), 2001.

[12] Timothy Chklovski and Patrick Pantel, *VerbOcean: Mining the Web for fine-grained semantic verb relations*, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP'04) (Barcelona, Spain), 2004, pp. 33–40.

[13] Andras Csomai, *Keywords in the mist: Automated keyword extraction for very large documents and back of the book indexing*, Ph.D. thesis, University of North Texas, 2008.

[14] Silviu Cucerzan and David Yarowsky, *Bootstrapping a multilingual part-of-speech tagger in one person-day*, Proceedings of the 6th Workshop on Computational Language Learning (CoNLL'02) (Taipei, Taiwan), Association for Computational Linguistics, 2002, pp. 132–138.

[15] Ido Dagan, Oren Glickman, and Bernardo Magnini, *The PASCAL recognising textual entailment challenge*, Proceedings of Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop (MLCW'05) (Southampton, UK), 2005, pp. 177–190.

[16] Bill Dolan, Chris Quirk, and Chris Brockett, *Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources*, Proceedings of the 20th International Conference on Computational Linguistics (Coling'04) (Geneva, Switzerland), 2004.

[17] Olivier Ferret, *Testing semantic similarity measures for extracting synonyms from a corpus*, Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10) (Valletta, Malta), 2010, pp. 3338–3343.

[18] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin, *Placing search in context: the concept revisited*, ACM Transactions on Information Systems 20 (2002), no. 1, 116–131.

[19] Evgeniy Gabrilovich and Shaul Markovitch, *Computing semantic relatedness using Wikipedia-based explicit semantic analysis*, Proceedings of the 20th AAAI International Conference on Artificial Intelligence (AAAI'07) (Hyderabad, India), 2007, pp. 1606–1611.

[20] William A. Gale, Kenneth W. Church, and David Yarowsky, *One sense per discourse*, Proceedings of the Workshop on Speech and Natural Language (HLT '91) (Harriman, New York, USA), Association for Computational Linguistics, 1992, pp. 233–237.

[21] Simon Garrod, *The role of different types of anaphor in the on-line resolution of sentences in a discourse*, Journal of Memory and Language 33 (1994), no. 1, 39–68.

[22] Jim Giles, *Internet encyclopaedias go head to head*, Nature 438 (2005), no. 7070, 900–901.

[23] Abby Goodrum, *Image information retrieval: An overview of current research*, Informing Science 3 (2000), no. 2, 63–67.

[24] Arthur C. Graesser, Keith K. Millis, and Rolf A. Zwaan, *Discourse comprehension*, Annual review of psychology 48 (1997), 163–189.

[25] Karl Haberlandt, *Methods in reading research*, pp. 1–31, Academic Press, Inc., San Diego, CA, US, 1994.

[26] Samer Hassan and Rada Mihalcea, *Cross-lingual semantic relatedness using encyclopedic knowledge*, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09) (Singapore), Association for Computational Linguistics, 2009, pp. 1192–1201.

[27] Nevin Heintze, *Scalable Document Fingerprinting*, Proceedings of 1996 USENIX Workshop on Electronic Commerce, USENIX, 1996, pp. 191–200.

[28] Graeme Hirst and David St-Onge, *Lexical Chains as representation of context for the detection and correction malapropisms*, The MIT Press, 1998.

[29] Timothy C. Hoad and Justin Zobel, *Methods for identifying versioned and plagiarized documents*, Journal of the American Society for Information Science and Technology 54 (2003), no. 3, 203–215.

[30] Thad Hughes and Daniel Ramage, *Lexical Semantic Relatedness with Random Graph Walks* , Processing (2007).

[31] Aminul Islam and Diana Inkpen, *Second order co-occurrence PMI for determining the semantic similarity of words*, Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 06) (Genoa, Italy), vol. 2, July 2006, pp. 1033–1038.

[32] ———, *Semantic text similarity using corpus-based word similarity and string similarity*, ACM Transactions on Knowledge Discovery from Data 2 (2008), no. 2, 1–25.

[33] ———, *Semantic similarity of short texts*, Current Issues in Linguistic Theory: Recent Advances in Natural Language Processing V (Nicolas Nicolov, Galia Angelova, and Ruslan Mitkov, eds.), John Benjamins Publishers, 2009, pp. 227–236.

[34] Mario Jarmasz, *Rogets thesaurus as a lexical resource for natural language processing (Ph.D. Thesis)*, Tech. report, University of Ottowa, 2003.

[35] Mario Jarmasz and Stan Szpakowics, *Rogets thesaurus and semantic similarity*, Proceedings of Recent Advances in Natural Language Processing (RANLP'03) (Borovetz, Bulgaria), 2003, pp. 111–120.

[36] Jay J Jiang, *Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy*, Computational Linguistics (1997), no. Rocling X.

[37] Thorsten Joachims, *A probabilistic analysis of the rocchio algorithm with tf.idf for text categorization*, Proceedings of the 14th International Conference on Machine Learning, Morgan Kaufmann Publishers Inc, 1997, pp. 143–151.

[38] Marcel Just and Patricia Carpenter, *A theory of reading: from eye fixation to comprehension*, Psychological Review 87 (1980), 329–354.

[39] Walter Kintsch, *The role of knowledge in discourse comprehension: A construction-integration model*, Psychological Review 95 (1988), no. 2, 163–182.

[40] Kazuaki Kishida, Kuang-hua Chen, Sukhoon Lee, Hsin-Hsi Chen, Noriko Kando, Kazuko Kuriyama, Sung Hyon Myaeng, and Koji Eguchi, *Cross-lingual information retrieval (CLIR) task at the NTCIR workshop 3*, ACM SIGIR Forum 38 (2004), no. 1, 17.

[41] Thomas K Landauer and Susan T Dumais, *A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.*, Psychological Review 104 (1997), no. 2, 211–240.

[42] Thomas K Landauer, Peter W Foltz, and Darrell Laham, *An introduction to Latent Semantic Analysis*, Discourse Processes (1998), no. 25, 259–284.

[43] Mirella Lapata and Regina Barzilay, *Automatic evaluation of text coherence: models and representations*, Proceedings of the 19th International Joint Conference on Artificial Intelligence (Edinburgh, Scotland), Morgan Kaufmann Publishers Inc., 2005, pp. 1085–1090.

[44] Claudia Leacock and Martin Chodorow, *Combining local context and WordNet similarity for word sense identification*, pp. 305–332, 1998.

[45] Michael D. Lee and Matthew Welsh, *An empirical evaluation of models of text document similarity*, Proceedings of the 27th annual meeting of the Cognitive Science Society (CogSci'05) (Stresa, Italy), 2005, pp. 1254–1259.

[46] Douglas B. Lenat and Edward A. Feigenbaum, *On the thresholds of knowledge*, Artificial Intelligence 47 (1991), no. 1-3, 185–250.

[47] Douglas B. Lenat and Ramanathan V. Guha, *Building large knowledge-based Systems; representation and inference in the Cyc project*, 1st ed., Addison-Wesley Longman Publishing Co., Inc., 1990.

[48] Chee Wee Leong and Rada Mihalcea, *Explorations in automatic image annotation using textual features*, Proceedings of the Third Linguistic Annotation Workshop (Suntec, Singapore), Association for Computational Linguistics, 2009, pp. 56–59.

[49] Michael Lesk, *Automatic sense disambiguation using machine readable dictionaries*, Proceedings of the 5th annual international conference on Systems documentation (SIGDOC'86) (Toronto, Ontario), ACM Press, 1986, pp. 24–26.

[50] Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Crockett Crockett, *Sentence similarity based on semantic nets and corpus statistics*, IEEE Transactions on Knowledge and Data Engineering 18 (2006), no. 8, 1138–1150.

[51] Chin-Yew Lin and Eduard Hovy, *Automatic evaluation of summaries using N-gram co-occurrence statistics*, Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03 (Edmonton, Canada), Association for Computational Linguistics, 2003, pp. 71–78.

[52] Dekang Lin, *An information-theoretic definition of similarity*, Proceedings of the Fifteenth International Conference on Machine Learning (ICML'10) (Madison, Wisconsin), 1998, pp. 296–304.

[53] Udi Manber, *Finding similar files in a large file system*, Proceedings of the USENIX Winter Technical Conference on USENIX Winter 1994 Technical Conference (San Francisco, California), USENIX Association, 1994, p. 2.

[54] Irina Matveeva, Gina-Anne Royer, and Levow Christiaan, *Term Representation with Generalized Latent Semantic Analysis*, Proceeding of the Recent Advances in Natural Language Processing (RANLP) (Borovets, Bulgaria), 2005.

[55] Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll, *Finding predominant word senses in untagged text*, Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - (ACL) (Barcelona, Spain), Association for Computational Linguistics, 2004, p. 279.

[56] Danielle S. Mcnamara, Eileen Kintsch, Nancy Butler Songer, and Walter Kintsch, *Are good texts always better? Text coherence, background knowledge, and levels of understanding in learning from text*, Cognition and Instruction 14 (1996), no. 1, 1–43.

[57] Donald Metzler, Yaniv Bernstein, W. Bruce Croft, Alistair Moffat, and Justin Zobel, *Similarity measures for tracking information flow*, Proceedings of the 14th ACM international conference on Information and knowledge management - CIKM '05 (2005), 517.

[58] Rada Mihalcea and Carlo Corley, Courtney Strapparava, *Corpus-based and knowledge-based measures of text semantic similarity*, Proceedings of the 21st National Conference on Artificial intelligence (Boston, Massachusetts), AAAI Press, 2006, pp. 775–780.

[59] Rada Mihalcea and Andras Csomai, *Wikify!: linking documents to encyclopedic knowledge*, Proceedings of the 16th ACM Conference on Information and Knowledge Management - (CIKM) (Lisbon, Portugal), ACM Press, 2007, pp. 233–242.

[60] George A. Miller, *WordNet: a Lexical database for English*, Communications of the Association for Computing Machinery 38 (1995), no. 11, 39–41.

[61] George A. Miller and Walter G. Charles, *Contextual correlates of semantic similarity*, Language and Cognitive Processes 6 (1991), no. 1, 1–28.

[62] Tom Mitchell, Terry Russell, Peter Broomhead, and Nicola Aldridge, *Towards robust computerised marking of free-text responses*, Proceedings of the 6th International Computer Assisted Assessment (CAA) Conference (Loughborough, UK), Loughborough University, 2002.

[63] Saif Mohammad, Iryna Gurevych, Graeme Hirst, and Torsten Zesch, *Cross-lingual distributional profiles of concepts for measuring semantic distance*, Proceeding of theProceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural (EMNLP-CoNLL) Language Learning (Prague, Czech Republic), ACL, 2007, pp. 571–580.

[64] Michael Mohler and Rada Mihalcea, *Text-to-text semantic similarity for automatic short answer grading*, Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09) (Stroudsburg, PA, USA), 2009, pp. 567–575.

[65] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-jing Zhu, *Bleu: a method for automatic evaluation of machine translation*, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (Philadelphia, PA), 2002, pp. 311–318.

[66] Siddharth Patwardhan and Ted Pedersen, *Using WordNet-based context vectors to estimate the semantic relatedness of concepts*, Proceeding of the Conference of the European Chapter of the Association for Computational Linguistics (EACL) (Trento, Italy), 2006, pp. 1–8.

[67] Viktor Pekar and Steffen Staab, *Word classification based on combined measures of distributional and semantic similarity*, Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - EACL '03 (2003), 147.

[68] Plato, *Allegory of the cave*, ch. Book VII, Athens, 380.

[69] Jay M. Ponte and Bruce W. Croft, *A Language modeling approach to information retrieval*, Proceedings of the Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval (Melbourne, Australia), 1998, pp. 275–281.

[70] Simone Paolo Ponzetto and Michael Strube, *Knowledge derived from wikipedia for computing semantic relatedness*, Journal of Artificial Intelligence Research 30 (2007), no. 1, 181–212.

[71] Adrian Popescu, Theodora Tsikrika, and Jana Kludas, *Overview of the Wikipedia Retrieval Task at ImageCLEF 2010*, CLEF 2010 Labs and Workshops, Notebook Papers (Padua, Italy), 2010.

[72] Stephen G. Pulman and Jana Z. Sukkarieh, *Automatic short answer marking*, Proceedings of the second workshop on Building Educational Applications Using Natural Language Processing (Ann Arbor, Michigan), Association for Computational Linguistics, 2005, pp. 9–16.

[73] Vilayanur S. Ramachandran, *The future of brain research*, 2003.

[74] Philip Resnik, *Using information content to evaluate semantic similarity in a taxonomy*, Proceedings of the 14th International Joint Conference on Artificial Intelligence (Montreal, Quebec, Canada), Morgan Kaufmann Publishers Inc., 1995, pp. 448–453.

[75] Joseph John Rocchio, *Relevance feedback in information retrieval*, pp. 313–323, Prentice Hall, Upper Saddle River, New Jersey, 1971.

[76] Herbert Rubenstein and John B. Goodenough, *Contextual correlates of synonymy*, Communications of the ACM 8 (1965), no. 10, 627–633.

[77] Oliver Sacks, *Musicophilia: Tales of music and the brain*, Knopf, 2007.

[78] Mehran Sahami and Timothy D. Heilman, *A web-based kernel function for measuring the similarity of short text snippets*, Proceedings of the 15th international conference on World Wide Web - WWW '06 (New York, New York, USA), ACM Press, 2006, p. 377.

[79] Gerard Salton and Christopher Buckley, *Term-weighting approaches in automatic text retrieval*, Information Processing & Management 24 (1988), no. 5, 513–523.

[80] Gerard Salton and Michael Lesk, *Computer evaluation of indexing and text processing*, Journal of the ACM 15 (1968), no. 1, 8–36.

[81] Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley, *Automatic text structuring and summarization*, Information Processing & Management 33 (1997), no. 2, 193–207.

[82] Hinrich Schütze, *Automatic Word Sense Discrimination*, Computational Linguistics 24 (1998), no. 1, 97–123.

[83] Narayanan Shivakumar and Hector Garcia-molina, *SCAM: A copy detection mechanism for digital documents*, Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries (TPDL'95) (Austin, Texas), 1995.

[84] John Sinclair, *Collins cobuild English dictionary for advanced learners*, 3rd ed., Harper Collins, 2001.

[85] Michael Strube and Simone Paolo Ponzetto, *WikiRelate! computing semantic relatedness using wikipedia*, Proceeding of the 21st AAAI Conference on Artificial Intelligence (AAAI'06) (Boston, Massachusetts), AAAI Press, 2006, pp. 1419–1424.

[86] Peter Turney, *Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL**, Lecture Notes in Computer Science (2001), no. September, 491–502.

[87] Teun A. van Dijk and Walter Kintsch, *Strategies of discourse comprehension*, Academic Press, Inc., New York, New York, USA, 1983.

[88] Peter Wiemer-Hastings, Katja Wiemer-Hastings, and Arthur C. Graesser, *Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis*, Proceeding of the Artificial Intelligence in Education (AIED'99) (Amsterdam), IOS Press, 1999, pp. 535–542.

[89] Zhibiao Wu and Martha Palmer, *Verbs semantics and lexical selection*, Proceedings of the 32nd annual meeting of the Association for Computational Linguistics (ACL'94) (Las Cruces, New Mexico), 1994, pp. 133–138.

[90] Wen-tau Yih and Christopher Meek, *Improving similarity measures for short segments of text*, Proceedings of the 22nd Conference on Artificial Intelligence (AAAI'07) (Vancouver, British Columbia, Canada), AAAI Press, 2007, pp. 1489–1494.

[91] Torsten Zesch, Christof Müller, and Iryna Gurevych, *Using Wiktionary for computing semantic relatedness*, Proceedings of the 23rd Conference on Articial Intelligence (AAAI'08) (Chicago, Illinois), AAAI Press, 2008, pp. 861–867.

[92] Rolf A. Zwaan, *Effect of genre expectations on text comprehension*, Journal of Experimental Psychology: Learning, Memory, and Cognition 20 (1994), 920–933.