Cross-lingual Semantic Relatedness Using Encyclopedic Knowledge

Samer Hassan and **Rada Mihalcea**
Department of Computer Science
University of North Texas
samer@unt.edu, rada@cs.unt.edu

Abstract

In this paper, we address the task of cross-lingual semantic relatedness. We introduce a method that relies on the information extracted from Wikipedia, by exploiting the interlanguage links available between Wikipedia versions in multiple languages. Through experiments performed on several language pairs, we show that the method performs well, with a performance comparable to monolingual measures of relatedness.

1 Motivation

Given the accelerated growth of the number of multilingual documents on the Web and elsewhere, the need for effective multilingual and cross-lingual text processing techniques is becoming increasingly important. In this paper, we address the task of *cross-lingual semantic relatedness*, and introduce a method that relies on Wikipedia in order to calculate the relatedness of words across languages. For instance, given the word *factory* in English and the word *lavoratore* in Italian (En. *worker*), the method can measure the relatedness of these two words despite the fact that they belong to two different languages.

Measures of cross-language relatedness are useful for a large number of applications, including cross-language information retrieval (Nie et al., 1999; Monz and Dorr, 2005), cross-language text classification (Gliozzo and Strapparava, 2006), lexical choice in machine translation (Och and Ney, 2000; Bangalore et al., 2007), induction of translation lexicons (Schafer and Yarowsky, 2002), cross-language annotation and resource projections to a second language (Riloff et al., 2002; Hwa et al., 2002; Mohammad et al., 2007).

The method we propose is based on a measure of closeness between concept vectors automatically built from Wikipedia, which are mapped via the Wikipedia interlanguage links. Unlike previous methods for cross-language mapping, which are typically limited by the availability of bilingual dictionaries or parallel texts, the method proposed in this paper can be used to measure the relatedness of word pairs in any of the 250 languages for which a Wikipedia version exists.

The paper is organized as follows. We first provide a brief overview of Wikipedia, followed by a description of the method to build concept vectors based on this encyclopedic resource. We then show how these concept vectors can be mapped across languages for a cross-lingual measure of word relatedness. Through evaluations run on six language pairs, connecting English, Spanish, Arabic and Romanian, we show that the method is effective at capturing the cross-lingual relatedness of words, with results comparable to the monolingual measures of relatedness.

2 Wikipedia

Wikipedia is a free online encyclopedia, representing the outcome of a continuous collaborative effort of a large number of volunteer contributors. Virtually any Internet user can create or edit a Wikipedia webpage, and this “freedom of contribution” has a positive impact on both the quantity (fast-growing number of articles) and the quality (potential errors are quickly corrected within the collaborative environment) of this online resource.

The basic entry in Wikipedia is an *article* (or *page*), which defines and describes an entity or an event, and consists of a hypertext document with hyperlinks to other pages within or outside Wikipedia. The role of the hyperlinks is to guide the reader to pages that provide additional information about the entities or events mentioned in an article. Articles are organized into *categories*, which in turn are organized into hierarchies. For instance, the article *automobile* is included in the category *vehicle*, which in turn has a parent cate-

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1192–1201,
Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

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

Table 1: Top ten largest Wikipedias

gory named *machine*, and so forth.

Each article in Wikipedia is uniquely referenced by an identifier, consisting of one or more words separated by spaces or underscores and occasionally a parenthetical explanation. For example, the article for *bar* with the meaning of *"counter for drinks"* has the unique identifier *bar (counter)*.

Wikipedia editions are available for more than 250 languages, with a number of entries varying from a few pages to two millions articles or more per language. Table 1 shows the ten largest Wikipedias (as of December 2008), along with the number of articles and approximate number of contributors.[1]

Relevant for the work described in this paper are the *interlanguage links*, which explicitly connect articles 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 to articles in other languages. The number of interlanguage links per article varies from an average of five in the English Wikipedia, to ten in the Spanish Wikipedia, and as many as 23 in the Arabic Wikipedia.

3 Concept Vector Representations using Explicit Semantic Analysis

To calculate the cross-lingual relatedness of two words, we measure the closeness of their concept vector representations, which are built from Wikipedia using explicit semantic analysis (ESA).

Encyclopedic knowledge is typically organized into concepts (or topics), each concept being further described using definitions, examples, and possibly links to other concepts. ESA (Gabrilovich and Markovitch, 2007) relies on the distribution of words inside the encyclopedic descriptions, and builds semantic representations for a given word in the form of a vector of the encyclopedic concepts in which the word appears. In this vector representation, each encyclopedic concept is assigned with a weight, calculated as the term frequency 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. We define $\vec{a}$ as the ESA concept vector of term a:

$$\vec{a} = \{w_{c_1}, w_{c_2} ... w_{c_n}\}, \qquad (1)$$

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

We use a revised version of the ESA algorithm. The original ESA semantic relatedness between the words in a given word pair $a - b$ is defined as the cosine similarity between their corresponding vectors:

$$Relatedness(a, b) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \left\|\vec{b}\right\|}. \qquad (2)$$

To illustrate, consider for example the construction of the ESA concept vector for the word *bird*. The top ten concepts containing this word, along with the associated weight (calculated using equation 7), are listed in table 2. Note that the the ESA vector considers all the possible senses of *bird*, including *Bird* as a surname as in e.g., *"Larry Bird."*

Weight	Wikipedia concept
51.4	Lists Of Birds By Region
44.8	Bird
40.3	British Birds Rarities Committee
32.8	Origin Of Birds
31.5	Ornithology
30.1	List Of Years In Birding And Ornithology
29.8	Bird Vocalization
27.4	Global Spread Of H5n1 In 2006
26.5	Larry Bird
22.3	Birdwatching

Table 2: Top ten Wikipedia concepts for the word "bird"

In our ESA implementation, we make three changes with respect to the original ESA algorithm. First, we replace the cosine similarity with

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

a Lesk-like metric (Lesk, 1986), which places less emphasis on the distributional differences between the vector weights and more emphasis on the overlap (mutual coverage) between the vector features, and thus it is likely to be more appropriate for the sparse ESA vectors, and for the possible asymmetry between languages. Let a and b be two terms with the corresponding ESA concept vectors $\vec{A}$ and $\vec{B}$ respectively. Let A and B represent the sets of concepts with a non-zero weight encountered in $\vec{A}$ and $\vec{B}$ respectively. The coverage of $\vec{A}$ by $\vec{B}$ is defined as:

$$G(\vec{B}|\vec{A}) = \sum_{i \in B} w_{a_i} \tag{3}$$

and similarly, the coverage of $\vec{B}$ by $\vec{A}$ is:

$$G(\vec{A}|\vec{B}) = \sum_{i \in A} w_{b_i} \tag{4}$$

where wa_i and wb_i represent the weight associated with concept c_i in vectors $\vec{A}$ and $\vec{B}$ respectively. By averaging these two asymmetric scores, we redefine the relatedness as:

$$Relatedness(a, b) = \frac{G(\vec{B}|\vec{A}) + G(\vec{A}|\vec{B})}{2} \tag{5}$$

Second, we refine the ESA weighting schema to account for the length of the articles describing the concept. Since some concepts have lengthy descriptions, they may be favored due to their high term frequencies when compared to more compact descriptions. To eliminate this bias, we calculate the weight associated with a concept c_i as follows:

$$w_{c_i} = tf_i \times log(M/\,|c_i|), \tag{6}$$

where tf_i represents the term frequency of the word a in concept c_i, M is a constant representing the maximum vocabulary size of Wikipedia concepts, and $|c_i|$ is the size of the vocabulary used in the description of concept c_i.

Finally, we use the Wikipedia category graph to promote category-type concepts in our feature vectors. This is done by scaling the concept's weight by the inverse of the distance d_i to the root category. The concepts that are not categories are treated as leaves, and therefore their weight is scaled down by the inverse of the maximum depth in the category graph. The resulting weighting scheme is:

$$w_{c_i} = tf_i \times log(M/\,|c_i|)/d_i \tag{7}$$

4 Cross-lingual Relatedness

We measure the relatedness of concepts in different languages by using their ESA concept vector representations in their own languages, along with the Wikipedia interlanguage links that connect articles written in a given language to their corresponding Wikipedia articles in other languages. For example, the English Wikipedia article *moon* contains interlanguage links to قمر in the Arabic Wikipedia, *luna* in the Spanish Wikipedia, and *lună* in the Romanian Wikipedia. The interlanguage links can map concepts across languages, and correspondingly map concept vector representations in different languages.

Formally, let C_x and C_y be the sets of all Wikipedia concepts in languages x and y, with corresponding translations in the y and x languages, respectively. If $tr_{xy}()$ is a translation function that maps a concept $c_i \in C_x$ into the concept $c_i^{'} \in C_y$ via the interlanguage links, we can write:

$$tr_{xy}(c_i) = c_i^{'}, \tag{8}$$

The projection of the ESA vector $\vec{t}$ from language x onto y can be written as:

$$tr_{xy}(\vec{t}) = \left\{ w_{tr_{xy}(c_1)} ... w_{tr_{xy}(c_n)} \right\}. \tag{9}$$

Using equations 5, 7, and 9, we can calculate the cross-lingual semantic relatedness between any two content terms a_x and b_y in given languages x and y as:

$$sim(a_x, b_y) = \frac{G(tr_{yx}(\vec{B})|\vec{A}) + G(\vec{A}|tr_{yx}(\vec{B}))}{2}. \tag{10}$$

Note that the weights assigned to Wikipedia concepts inside the concept vectors are language specific. That is, two Wikipedia concepts from different languages, mapped via an interlanguage link, can, and often do have different weights.

Intuitively, the relation described by the interlanguage links should be reflective and transitive. However, due to Wikipedia's editorial policy, which accredits users with the responsibility of maintaining the articles, these properties are not always met. Table 3 shows real cases where the transitive and the reflective properties fail due to missing interlanguage links.

Relation	Exists
Reflectivity	
Kafr-El-Dawwar Battle(en) ↦ معركة كفر الدوَار(ar)	Yes
معركة كفر الدوَار(ar) ↦ Kafr-El-Dawwar Battle(en)	No
Transitivity	
Intifada(en) ↦ Intifada(es)	Yes
Intifada(es) ↦ انتفَاضة(ar)	Yes
Intifada(en) ↦ انتفَاضة(ar)	No

Table 3: Reflectivity and transitivity in Wikipedia

We solve this problem by iterating over the translation tables and extracting all the missing links by enforcing the reflectivity and the transitivity properties. Table 4 shows the initial number of interlanguage links and the discovered links for the four languages used in our experiments. The table also shows the coverage of the interlanguage links, measured as the ratio between the total number of interlanguage links (initial plus discovered) originating in the source language towards the target language, divided by the total number of articles in the source language.

	Interlanguage links		
Language pair	Initial	Discov.	Cover.
English → Spanish	293,957	12,659	0.14
English → Romanian	86,719	4,641	0.04
English → Arabic	56,233	3,916	0.03
Spanish → English	294,266	7,328	0.58
Spanish → Romanian	39,830	3,281	0.08
Spanish → Arabic	33,889	3,319	0.07
Romanian → English	75,685	6,783	0.46
Romanian → Spanish	36,002	3,546	0.22
Romanian → Arabic	15,777	1,698	0.10
Arabic → English	46,072	3,170	0.33
Arabic → Spanish	28,142	3,109	0.21
Arabic → Romanian	15,965	1,970	0.12

Table 4: Interlanguage links (initial and discovered) and their coverage in Wikipedia versions in four languages.

5 Experiments and Evaluations

We run our experiments on four languages: English, Spanish, Romanian and Arabic. For each of these languages, we use a Wikipedia download from October 2008. The articles were preprocessed using Wikipedia Miner (Milne, 2007) to extract structural information such as generality, and interlanguage links. Furthermore, articles were also processed to remove numerical content, as well as any characters not included in the given language's alphabet. The content words are stemmed, and words shorter than three characters are removed (a heuristic which we use as an approximation for stopword removal). Table 5 shows the number of articles in each Wikipedia version and the size of their vocabularies, as obtained after the pre-processing step.

	Articles	Vocabulary
English	2,221,980	1,231,609
Spanish	520,154	406,134
Arabic	149,340	216,317
Romanian	179,440	623,358

Table 5: Number of articles and size of vocabulary for the four Wikipedia versions

After pre-processing, the articles are indexed to generate the ESA concept vectors. From each Wikipedia version, we also extract other features including article titles, interlanguage links, and Wikipedia category graphs. The interlanguage links are further processed to recover any missing links, as described in the previous section.

5.1 Data

For the evaluation, we build several cross-lingual datasets based on the standard Miller-Charles (Miller and Charles, 1998) and WordSimilarity-353 (Finkelstein et al., 2001) English word relatedness datasets.

The Miller-Charles dataset (Miller and Charles, 1998) consists of 30-word pairs ranging from synonymy pairs (e.g., *car - automobile*) to completely unrelated terms (e.g., *noon - string*). The relatedness of each word pair was rated by 38 human subjects, using a scale from 0 (not-related) to 4 (perfect synonymy). The dataset is available only in English and has been widely used in previous semantic relatedness evaluations (e.g., (Resnik, 1995; Hughes and Ramage, 2007; Zesch et al., 2008)).

The WordSimilarity-353 dataset (also known as Finkelstein-353) (Finkelstein et al., 2001) 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 data set. Unlike the Miller-Charles data set, which consists only of

	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	maşfină - automobil	frate - călugăr

Table 6: Word pair translation examples

single words, the WordSimilarity-353 set also features phrases (e.g., “*Wednesday news*”), therefore posing an additional degree of difficulty for a relatedness metric applied on this data.

Native speakers of Spanish, Romanian and Arabic, who were also highly proficient in English, were asked to translate the words in the two data sets. The annotators were provided one word pair at a time, and asked to provide the appropriate translation for each word while taking into account their relatedness within the word pair. The relatedness was meant as a hint to disambiguate the words, when multiple translations were possible.

The annotators were also 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*, annotators were allowed to use دينَار[2] as a translation for *buck*.

To test the ability of the bilingual judges to provide correct translations by using this annotation setting, we carried out the following experiment. We collected Spanish translations from five different human judges, which were then merged into a single selection based on the annotators’ 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.

Subsequently, five additional human experts rescored the word-pair Spanish translations by using the same scale that was used in the construction of the English data set. 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 word relatedness.

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 setting was effective in pinpointing the correct translation for each word, even in the case of ambiguous words.

Motivated by the validation of the annotation setting obtained for Spanish, we used only one human annotator to collect the translations for Arabic and Romanian. Table 6 shows examples of translations in the three languages for three word pairs from our data sets.

Using these translations, we create six cross-lingual data sets, one for each possible language pair (English-Spanish, English-Arabic, English-Romanian, Spanish-Arabic, Spanish-Romanian, Arabic-Romanian). Given a source-target language pair, a data set is created by first using the source language for the first word and the target language for the second word, and then reversing the order, i.e., using the source language for the second word and the target language for the first word. The size of the data sets is thus doubled in this way (e.g., the 30 word pairs in the English Miller-Charles set are transformed into 60 word pairs in the English-Spanish Miller-Charles set).

5.2 Results

We evaluate the cross-lingual measure of relatedness on each of the six language pairs. For comparison purposes, we also evaluate the monolingual relatedness on the four languages.

For the evaluation, we use the Pearson (r) and Spearman (ρ) correlation coefficients, which are the standard metrics used in the past for the evaluation of semantic relatedness (Finkelstein et

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

al., 2001; Zesch et al., 2008; Gabrilovich and Markovitch, 2007). While the Pearson correlation is highly dependent on the linear relationship between the distributions in question, Spearman mainly emphasizes the ability of the distributions to maintain their relative ranking.

Tables 7 and 8 show the results of the evaluations of the cross-lingual relatedness, when using an ESA concept vector with a size of maximum 10,000 concepts.[3]

	English	Spanish	Arabic	Romanian
	Miller-Charles			
English	0.58	0.43	0.32	0.50
Spanish		0.44	0.20	0.38
Arabic			0.36	0.32
Romanian				0.58
	WordSimilarity-353			
English	0.55	0.32	0.31	0.29
Spanish		0.45	0.32	0.28
Arabic			0.28	0.25
Romanian				0.30

Table 7: Pearson correlation for cross-lingual relatedness on the Miller-Charles and WordSimilarity-353 data sets

	English	Spanish	Arabic	Romanian
	Miller-Charles			
English	0.75	0.56	0.27	0.55
Spanish		0.64	0.17	0.32
Arabic			0.33	0.21
Romanian				0.61
	WordSimilarity-353			
English	0.71	0.55	0.35	0.38
Spanish		0.50	0.29	0.30
Arabic			0.26	0.20
Romanian				0.28

Table 8: Spearman correlation for cross-lingual relatedness on the Miller-Charles and WordSimilarity-353 data sets

As a validation of our ESA implementation, we compared the results obtained for the monolingual English relatedness with other results reported in the past for the same data sets. Gabrilovich and Markovitch (2007) reported a Spearman correlation of 0.72 for the Miller-Charles data set and 0.75 for the WordSimilarity-353 data set, respectively. Zesch et al. (2008) reported a Spearman correlation of 0.67 for the Miller-Charles set. These values are comparable to the Spearman correlation scores obtained in our experiments for the English data sets (see Table 8), with a fairly large improvement obtained on the Miller-Charles data set when using our implementation.

6 Discussion

Overall, our method succeeds in capturing the cross-lingual semantic relatedness between words. As a point of comparison, one can use the monolingual measures of relatedness as reflected by the diagonals in Tables 7 and 8.

Looking at the monolingual evaluations, the results seem to be correlated with the Wikipedia size for the corresponding language, with the English measure scoring the highest. These results are not surprising, given the direct relation between the Wikipedia size and the sparseness of the ESA concept vectors. A similar trend is observed for the cross-lingual relatedness, with higher results obtained for the languages with large Wikipedia versions (e.g., English-Spanish), and lower results for the languages with a smaller size Wikipedia (e.g., Arabic-Spanish).

For comparison, we ran two additional experiments. In the first experiment, we compared the coverage of our cross-lingual relatedness method to a direct use of the translation links available in Wikipedia. The cross-lingual relatedness is turned into a monolingual relatedness by using the interlanguage Wikipedia links to translate the first of the two words in a cross-lingual pair into the language of the second word in the pair.[4] From the total of 433 word pairs available in the two data sets, this method can produce translations for an average of 103 word pairs per language pair. This means that the direct Wikipedia interlanguage links allow the cross-lingual relatedness measure to be transformed into a monolingual relatedness in about 24% of the cases, which is a low coverage compared to the full coverage that can be obtained with our cross-lingual method of relatedness.

In an attempt to raise the coverage of the translation, we ran a second experiment where we used a state-of-the-art translation engine to translate the first word in a pair into the language of the sec-

[3] The concepts are selected in reversed order of their weight inside the vector in the respective language. Note that the cross-lingual mapping between the concepts in the ESA vectors is done after the selection of the top 10,000 concepts in each language.

[4] We use all the interlanguage links obtained by combining the initial and the discovered links, as described in Section 4.

ond word in the pair. We use Google Translate, which is a statistical machine translation engine that relies on large parallel corpora, to find the most likely translation for a given word. Unlike the previous experiment, this time we can achieve full translation coverage, and thus we are able to produce data sets of equal size that can be used for a comparison between relatedness measures. Specifically, using the translation produced by the machine translation engine for the first word in a pair, we calculate the relatedness within the space of the language of the second word using a monolingual ESA also based on Wikipedia. The results obtained with this method are compared against the results obtained with our cross-lingual ESA relatedness.

Using a Pearson correlation, our cross-lingual relatedness method achieves an average score across all six language pairs of 0.36 for the Miller-Charles data set and 0.30 for the WordSimilarity-353 data set,[5] which is higher than the 0.33 and 0.28 scores achieved for the same data sets when using a translation obtained with Google Translate followed by a monolingual measure of relatedness. These results are encouraging, also given that the translation-based method is limited to those language pairs for which a translation engine exists (e.g., Google Translate covers 40 languages), whereas our method can be applied to any language pair from the set of 250 languages for which a Wikipedia version exists.

To gain further insights, we also determined the impact of the vector length in the ESA concept vector representation, by calculating the Pearson correlation for vectors of different lengths. Figures 1 and 2 show the Pearson score as a function of the vector length for the Miller-Charles and WordSimilarity-353 data sets. The plots show that the cross-lingual measure of relatedness is not significantly affected by the reduction or increase of the vector length. Thus, the use of vectors of length 10,000 (as used in most of our experiments) appears as a reasonable tradeoff between accuracy and performance.

Furthermore, by comparing the performance of the proposed Lesk-like model to the traditional cosine-similarity (Figures 3 and 4), we note that the Lesk-like model outperforms the cosine model on most language pairs. We believe that this is due to the stricter correlation conditions imposed by the cosine-metric in such sparse vector-based representations, as compared to the more relaxed hypothesis used by the Lesk model.

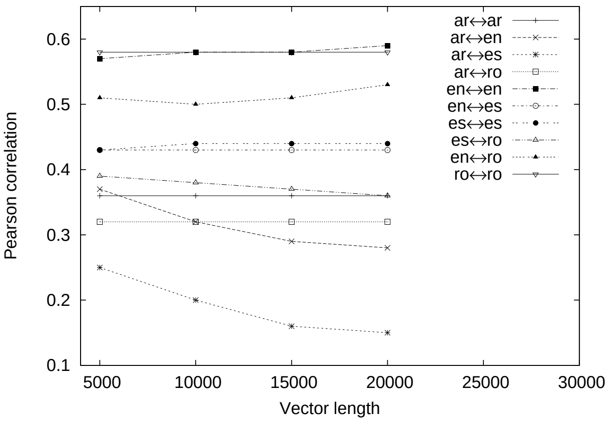

Figure 1: Pearson correlation vs. ESA vector length on the Miller-Charles data set

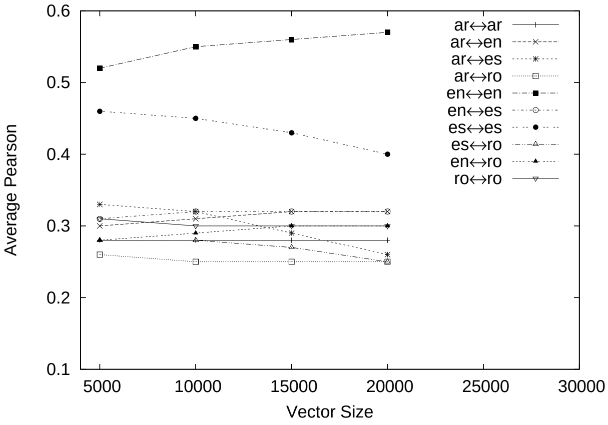

Figure 2: Pearson correlation vs. ESA vector length on the WordSimilarity-353 data set

Finally, we also looked at the relation between the number of interlanguage links found for the concepts in a vector and the length of the vector. Figures 5 and 6 display the average number of interlanguage links as a function of the concept vector length.

By analyzing the effect of the average number of interlanguage links found per word in the given datasets (Figures 5 and 6), we notice that these links increase proportionally with the vector size, as expected. However, this increase does not lead to any significant improvements in accuracy (Figures 1 and 2). This implies that while the presence of interlanguage links is a prerequisite for the mea-

[5] This average considers all the cross-lingual relatedness scores listed in Table 7; it does not include the monolingual scores listed on the table diagonal.

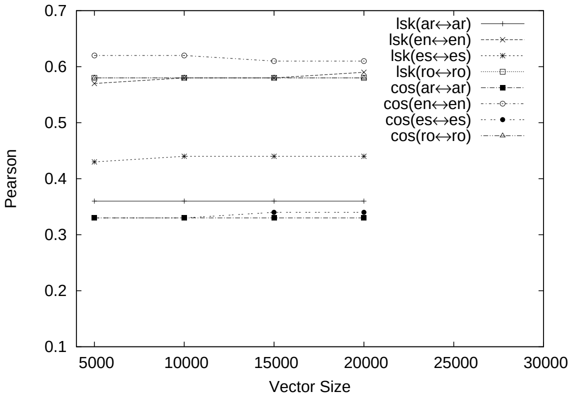

Figure 3: Lesk vs. cosine similarity for the Miller-Charles data set

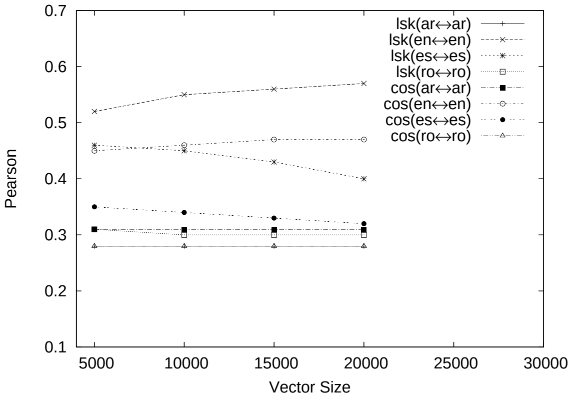

Figure 4: Lesk vs. cosine similarity for the WordSimilarity-353 data set

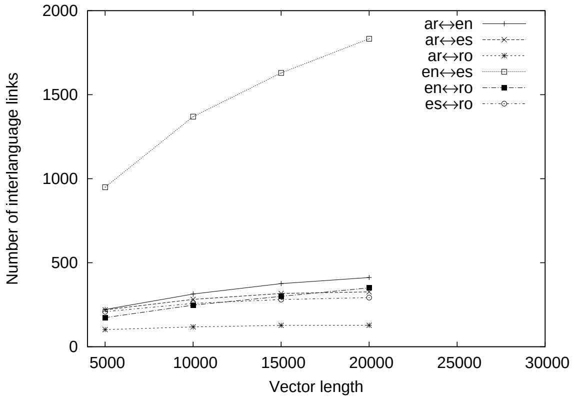

Figure 5: Number of interlanguage links vs. vector length for the Miller-Charles data set

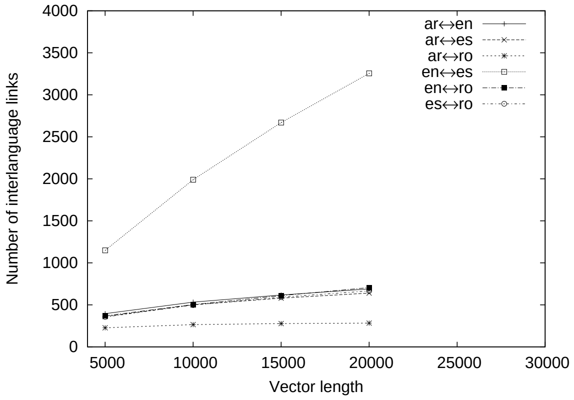

Figure 6: Number of interlanguage links vs. vector length for the WordSimilarity-353 data set

sure of relatedness,[6] their effect is only significant for the top ranked concepts in a vector. Therefore, increasing the vectors size to maximize the matching of the projected dimensions does not necessarily lead to accuracy improvements.

7 Related Work

Measures of word relatedness were found useful in a large number of natural language processing applications, including word sense disambiguation (Patwardhan et al., 2003), synonym identification (Turney, 2001), automated essay scoring (Foltz et al., 1999), malapropism detection (Budanitsky and Hirst, 2001), coreference resolution (Strube and Ponzetto, 2006), and others. Most of the work to date has focused on measures of word relatedness for English, by using methods applied on knowledge bases (Lesk, 1986; Wu and Palmer, 1994; Resnik, 1995; Jiang and Conrath, 1997; Hughes and Ramage, 2007) or on large corpora (Salton et al., 1997; Landauer et al., 1998; Turney, 2001; Gabrilovich and Markovitch, 2007).

Although to a lesser extent, measures of word relatedness have also been applied on other languages, including German (Zesch et al., 2007; Zesch et al., 2008; Mohammad et al., 2007), Chinese (Wang et al., 2008), Dutch (Heylen et al., 2008) and others. Moreover, assuming resources similar to those available for English, e.g., WordNet structures or large corpora, the measures of relatedness developed for English can be in principle applied to other languages as well.

All these methods proposed in the past have been concerned with *monolingual* word relatedness calculated within the boundaries of one language, as opposed to *cross-lingual* relatedness, which is the focus of our work.

The research area closest to the task of cross-

[6] Two languages with no interlanguage links between them will lead to a relatedness score of zero for any word pair across these languages, no matter how strongly related the words are.

lingual relatedness is perhaps cross-language information retrieval, which is concerned with matching queries posed in one language to document collections in a second language. Note however that most of the approaches to date for cross-language information retrieval have been based on direct translations obtained for words in the query or in the documents, by using bilingual dictionaries (Monz and Dorr, 2005) or parallel corpora (Nie et al., 1999). Such explicit translations can identify a direct correspondence between words in two languages (e.g., they will find that *fabbrica* (It.) and *factory* (En.) are translations of each other), but will not capture similarities of a different degree (e.g., they will not find that *lavoratore* (It.; worker in En.) is similar to *factory* (En.).

Also related are the areas of word alignment for machine translation (Och and Ney, 2000), induction of translation lexicons (Schafer and Yarowsky, 2002), and cross-language annotation projections to a second language (Riloff et al., 2002; Hwa et al., 2002; Mohammad et al., 2007). As with cross-language information retrieval, these areas have primarily considered direct translations between words, rather than an entire spectrum of relatedness, as we do in our work.

8 Conclusions

In this paper, we addressed the problem of cross-lingual semantic relatedness, which is a core task for a number of applications, including cross-language information retrieval, cross-language text classification, lexical choice for machine translation, cross-language projections of resources and annotations, and others.

We introduced a method based on concept vectors built from Wikipedia, which are mapped across the interlanguage links available between Wikipedia versions in multiple languages. Experiments performed on six language pairs, connecting English, Spanish, Arabic and Romanian, showed that the method is effective at capturing the cross-lingual relatedness of words. The method was shown to be competitive when compared to methods based on a translation using the direct Wikipedia links or using a statistical translation engine. Moreover, our method has wide applicability across languages, as it can be used for any language pair from the set of 250 languages for which a Wikipedia version exists.

The cross-lingual data sets introduced in this paper can be downloaded from http://lit.csci.unt.edu/index.php/Downloads.

Acknowledgments

The authors are grateful to Carmen Banea for her help with the construction of the data sets. This material is based in part upon work supported by the National Science Foundation CAREER award #0747340. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

S. Bangalore, P. Haffner, and S. Kanthak. 2007. Statistical machine translation through global lexical selection and sentence reconstruction. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic.

A. Budanitsky and G. Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh.

L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. 2001. Placing search in context: the concept revisited. In *WWW*, pages 406–414.

P. Foltz, D. Laham, and T. Landauer. 1999. Automated essay scoring: Applications to educational technology. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, Chesapeake, Virginia.

E. Gabrilovich and S. Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1606–1611.

A. Gliozzo and C. Strapparava. 2006. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *Proceedings of the Conference of the Association for Computational Linguistics*, Sydney, Australia.

K. Heylen, Y. Peirsman, D. Geeraerts, and D. Speelman. 2008. Modelling word similarity: an evaluation of automatic synonymy extraction algorithms. In *Proceedings of the Sixth International Language Resources and Evaluation*, Marrakech, Morocco.

T. Hughes and D. Ramage. 2007. Lexical semantic knowledge with random graph walks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Prague, Czech Republic.

R. Hwa, P. Resnik, A. Weinberg, and O. Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002),* Philadelphia, July.

J. Jiang and D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, Taiwan.

T. K. Landauer, P. Foltz, and D. Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25.

M.E. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the SIGDOC Conference 1986*, Toronto, June.

G. Miller and W. Charles. 1998. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1).

D. Milne. 2007. Computing semantic relatedness using wikipedia link structure. In European Language Resources Association (ELRA), editor, *In Proceedings of the New Zealand Computer Science Research Student Conference (NZCSRSC 2007)*, New Zealand.

S. Mohammad, I. Gurevych, G. Hirst, and T. Zesch. 2007. Cross-lingual distributional profiles of concepts for measuring semantic distance. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL-2007)*, Prague, Czech Republic.

C. Monz and B.J. Dorr. 2005. Iterative translation disambiguation for cross-language information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil.

J.-Y. Nie, M. Simard, P. Isabelle, and R. Durand. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*.

F. Och and H. Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrucken, Germany, August.

S. Patwardhan, S. Banerjee, and T. Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, February.

P. Resnik. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada.

E. Riloff, C. Schafer, and D. Yarowsky. 2002. Inducing information extraction systems for new languages via cross-language projection. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan, August.

G. Salton, A. Wong, and C.S. Yang. 1997. A vector space model for automatic indexing. In *Readings in Information Retrieval*, pages 273–280. Morgan Kaufmann Publishers, San Francisco, CA.

C. Schafer and D. Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL 2003)*, Taipei, Taiwan, August.

M. Strube and S. P. Ponzetto. 2006. Wikirelate! computing semantic relatedeness using Wikipedia. In *Proceedings of the American Association for Artificial Intelligence*, Boston, MA.

P. Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, Freiburg, Germany.

X. Wang, S. Ju, and S. Wu. 2008. A survey of chinese text similarity computation. In *Proceedings of the Asia Information Retrieval Symposium*, Harbin, China.

Z. Wu and M. Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico.

T. Zesch, I. Gurevych, and M. Mühlhäuser. 2007. Comparing Wikipedia and German Wordnet by Evaluating Semantic Relatedness on Multiple Datasets. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

T. Zesch, C. Müller, and I. Gurevych. 2008. Using Wiktionary for Computing Semantic Relatedness. In *Proceedings of the American Association for Artificial Intelligence*, Chicago.