Measuring Semantic Relatedness using Multilingual Representations

Samer Hassan
University of North Texas
Denton, TX
samer@unt.edu

Carmen Banea
University of North Texas
Denton, TX
carmenbanea@my.unt.edu

Rada Mihalcea
University of North Texas
Denton, TX
rada@cs.unt.edu

Abstract

This paper explores the hypothesis that semantic relatedness may be more reliably inferred by using a multilingual space, as compared to the typical monolingual representation. Through evaluations using several state-of-the-art semantic relatedness systems, applied on standard datasets, we show that a multilingual approach is better suited for this task, and leads to improvements of up to 47% with respect to the monolingual baseline.

1 Introduction

Semantic relatedness is the task of quantifying the strength of the semantic connection between textual units, be they words, sentences, or documents. For instance, one may want to determine how semantically related are two words such as car and $automobile$, or two pieces of text such as *I love animals* and *I own a pet.* 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 (Ponte and Croft, 1998), query reformulation (Metzler et al., 2007; Yih and Meek, 2007; Sahami and Heilman, 2006; Broder et al., 2008), image retrieval (Leong and Mihalcea, 2009; Goodrum, 2000), plagiarism detection (Hoad and Zobel, 2003; Shivakumar and Garcia-Molina, 1995; Broder et al., 1997; Heintze, 1996; Brin et al., 1995; Manber, 1994), information flow (Metzler et al., 2005), sponsored search (Broder et al., 2008), short answer grading (Mohler and Mihalcea, 2009a; Pulman and Sukkarieh, 2005; Mitchell et al., 2002), and textual entailment (Dagan et al., 2005).

The typical approach to semantic relatedness is to either measure the distance between the constituent words by using a knowledge base such as WordNet or Roget (e.g., (Leacock and Chodorow, 1998; Lesk, 1986; Jarmasz and Szpakowicz, 2003; Pedersen et al., 2004)), or to calculate the similarity between the word distributions in very large corpora (e.g., (Landauer et al., 1991; Lin, 1998; Gabrilovich and Markovitch, 2007)). With almost no exception, these methods have been applied on one language at a time – English, most of the time, although measures of relatedness have also been explored on languages such as German (Zesch et al., 2007), Chinese (Li et al., 2005), Japanese (Kazama et al., 2010), and others.

In this paper, we take a step further and explore a joint multilingual semantic relatedness metric, which aggregates semantic relatedness scores measured on several different languages. Specifically, in our method, in order to measure the relatedness of two textual units, we first determine their relatedness in multiple languages, and consequently infer a final relatedness score by averaging the scores calculated in the individual languages.

Our hypothesis is that a multilingual representation can enrich the relatedness space and address relevant issues such as *polysemy* (i.e., find that two occurrences of the same word in language L1 represent two different meanings because of different translations in language L2) and *synonymy* (i.e., find that two words in language L1 are related because they have the same translation in language L2). We show that by measuring relatedness in a multilingual space, we are able to improve over a traditional relatedness measure that relies exclusively on a monolingual representation.

Through experiments using several state-of-the-art measures of relatedness, applied on a multilingual space including English, Arabic, Spanish, and Romanian, we aim to answer the following research

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 20–29,
Montréal, Canada, June 7-8, 2012.

questions: (1) Does the task of semantic relatedness benefit from a multilingual representation, as compared to a monolingual one? (2) Does the translation quality affect the results? and (3) Do the findings hold for different relatedness datasets?

The paper is organized as follows. First, we overview related work on word and text relatedness, and on multilingual natural language processing. We then briefly describe three corpus-based measures of relatedness, and present several word and text datasets that have been used in the past to evaluate relatedness. We then present evaluations and experiments addressing each of the three research questions, and discuss our findings.

2 Related Work

Semantic relatedness. The approaches for semantic relatedness that have been considered to date can be grouped into knowledge-based and corpus-based. Knowledge-based methods derive a measure of relatedness by utilizing lexical resources and ontologies such as WordNet (Miller, 1995) to measure definitional overlap (Lesk, 1986), term distance within a graphical taxonomy (Leacock and Chodorow, 1998), term depth in the taxonomy as a measure of specificity (Wu and Palmer, 1994), and others. The application of such measures to a language other than English requires the availability of the lexical resource in that language; furthermore, even though taxonomies such as WordNet (Miller, 1995) are available in a number of languages[1], their coverage is still limited, and often times they are not publicly available. For these reasons, in multilingual settings, these measures often become untractable.

On the other side, corpus-based measures such as Latent Semantic Analysis (LSA) (Landauer et al., 1991), Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007), Salient Semantic Analysis (SSA) (Hassan and Mihalcea, 2011), Pointwise Mutual Information (PMI) (Church and Hanks, 1990), PMI-IR (Turney, 2001), Second Order PMI (Islam and Inkpen, 2006), Hyperspace Analogues to Language (HAL) (Burgess et al., 1998) and distributional similarity (Lin, 1998) employ probabilistic approaches to decode the semantics of words. They consist of unsupervised methods that utilize the contextual information and patterns observed in raw text to build semantic profiles of words, and thus they can be easily transferred to a new language provided that a large corpus in that language is available.

Multilingual natural language processing. Also relevant is the work done on multilingual text processing, which attempts to improve the performance of different natural language processing tasks by integrating information drawn from multiple languages. For instance, (Cohn and Lapata, 2007) explore the use of triangulation for machine translation, where multiple translation models are learned using multilingual parallel corpora. The model was found especially beneficial for languages where the training dataset was small, thus suggesting that this method may be particularly useful for languages with scarce resources. (Davidov and Rappoport, 2009) experiment with the use of multiple languages to enhance an existing lexicon. In their experiments, using three source languages and 45 intermediate languages, they find that the multilingual resources can lead to significant improvements in concept expansion. (Banea et al., 2010) explore the use of parallel multilingual corpora to improve subjectivity classification in a target language, finding that the use of multilingual representations for subjectivity analysis improves over the monolingual classifiers. Similarly, (Banea and Mihalcea, 2011) investigate the use of multilingual contexts for word sense disambiguation. By leveraging on the translations of the annotated contexts in multiple languages, a multilingual thematic space emerges that better disambiguates target words.

Finally, there are two lines of work that explore semantic distances in a multilingual space. First, (Besançon and Rajman, 2002) examine the notion that the distances between document vectors within a language correlate with the distances between their corresponding vectors in a parallel corpus. These findings provide clues about the possibility of reliable semantic knowledge transfer across language boundaries. Second, (Hassan and Mihalcea, 2009) propose a framework to compute semantic relatedness between two words in different languages, by considering Wikipedia articles in multiple languages. The method differs from the one proposed here, as we aggregate relatedness over monolingual spaces rather than measuring cross-lingual relatedness, and we do not specifically use the inter-wiki links between Wikipedia pages.

[1] http://www.illc.uva.nl/EuroWordNet/

3 Measures of Text Relatedness

In this work, we focus on corpus-based metrics because of their unsupervised nature, their flexibility, scalability, and portability to different languages. Specifically, we utilize three popular models, LSA (Landauer et al., 1991), ESA (Gabrilovich and Markovitch, 2007), and SSA (Hassan and Mihalcea, 2011). In these models, the semantic profile of a word is expressed in terms of the explicit (ESA), implicit (LSA), or salient (SSA) concepts. All three models are trained on the Wikipedia 2010 corpora corresponding to the four languages of interest (English, Arabic, Spanish, Romanian).

Explicit Semantic Analysis. ESA (Gabrilovich and Markovitch, 2007) uses encyclopedic knowledge in an information retrieval framework to generate a semantic interpretation of words. Since encyclopedic knowledge is typically organized into concepts (or topics), each concept is described using definitions and examples. ESA relies on the distribution of words inside the encyclopedic descriptions. It builds semantic representations for a given word using a word-document association, where each document represents a Wikipedia article. In this vector representation, the semantic interpretation of a text can be modeled as an aggregation of the semantic vectors of its individual words.

Latent Semantic Analysis. In LSA (Landauer et al., 1991), term-context associations are captured by means of a dimensionality reduction operated by a singular value decomposition (SVD) on the term-by-context matrix **T**, where the matrix is induced from a large corpus. This reduction entails the abstraction of meaning by collapsing similar contexts and discounting noisy and irrelevant ones, hence transforming the real world term-context space into a word-latent-concept space which achieves a much deeper and concrete semantic representation of words.

Salient Semantic Analysis. SSA (Hassan and Mihalcea, 2011) incorporates a similar semantic abstraction and interpretation of words, by using salient concepts gathered from encyclopedic knowledge, where a concept is defined as an unambiguous word or phrase with a concrete meaning, which can afford an encyclopedic definition. The links available between Wikipedia articles, obtained either through manual annotation by the Wikipedia users or using an automatic annotation process, are regarded as clues or salient features within the text that help define and disambiguate its context. This method seeks to determine the semantic relatedness of words by measuring the distance between their concept-based profiles, where a profile consists of co-occurring salient concepts found within a given window size in a very large corpus.

4 Datasets

To evaluate the representation strength of a multilingual semantic relatedness model we employ several standard word-to-word and text-to-text datasets. For each of these datasets, we make use of their representation in the four languages of interest.

4.1 Word Relatedness

We construct our multilingual word-to-word datasets building upon three word relatedness datasets that have been widely used in the past.

Rubenstein and Goodenough (Rubenstein and Goodenough, 1965) (**RG65**) 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 (unrelated) to 4 (synonyms).

Miller-Charles (Miller and Charles, 1991) (**MC30**) is a subset of $RG65$, consisting of 30 word pairs annotated for relatedness by 38 human subjects, using the same 0 to 4 scale.

WordSimilarity-353 (Finkelstein et al., 2001) (**WS353**), also known as Finkelstein-353, consists of 353 word pairs annotated by 13 human experts, on a scale from 0 (unrelated) to 10 (synonyms). While containing the $MC30$ set, it poses an additional degree of difficulty by also including phrases (e.g., *"Wednesday news"*), proper names and technical terms.

To enable a multilingual representation, we use the multilingual datasets introduced by (Hassan and Mihalcea, 2009), which are based upon $MC30$ and $WS353$. These multilingual datasets are built using manual translations, following the same guidelines adopted for the generation and the annotation of their original English counterparts. These manually translated collections, available in Arabic, Spanish, and Romanian, allow us to infer an upper bound for the multilingual semantic relatedness model.

Moreover, in order to provide a more realistic scenario, where manual translations are not available, we also create multilingual datasets by automatically translating the three English datasets into

Arabic, Spanish and Romanian.[2] Similar to how the manually translated datasets were created by providing the bilingual speakers with one word pair at a time, for the automatic translation each word pair is processed as a single query to the translation engine. Thus, the co-occurrence metrics derived from large corpora are able to play a role in providing a disambiguated translation instead of defaulting to the most frequently used sense if the words were to be processed individually. This allows for the embedded word pair relatedness to be transferred to other languages as well.

4.2 Text Relatedness

We use three standard text-to-text datasets.

Lee50 (Lee and Welsh, 2005) 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 (Lee and Welsh, 2005), the evaluations are carried out on only 1225 document pairs after ignoring duplicates.

Li30 (Li et al., 2006) is a sentence pair similarity dataset obtained by replacing each of the $RG65$ word-pairs with their respective definitions extracted from the Collins Cobuild dictionary (Sinclair, 2001). Each sentence pair was scored between 0 (unrelated) to 4 (alike) by 32 native English speakers, and their annotations were averaged. Due to the 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 relatedness distribution.

AG400 (Mohler and Mihalcea, 2009b) 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. We employ the version proposed by (Hassan and Mihalcea, 2011) which consists of 400 student answers along with the corresponding questions and correct instructor answers. Each student answer was graded by two judges on a scale from 0 (completely wrong) to 5 (perfect answer). The correlation between human judges was measured at 0.64.

First, we construct a multilingual, manually translated text-to-text relatedness dataset based on the standard $Li30$ corpus.[3] 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. They were presented with one sentence at a time, and asked to provide the appropriate translation into their native language. Since we had five Spanish, two Arabic, and two Romanian translators, 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, we asked three different Spanish human experts 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 to 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-pairs. As was the case for the manually constructed word-to-word datasets previously described, the metrics obtained on the manually translated $Li30$ dataset will also act as an upper bound for the text-to-text evaluations.

Finally, for a more sensible scenario where the text fragments do not require manual translations in order to compute their semantic relatedness, we create a multilingual version of the three English datasets by employing statistical machine translation to translate the texts into the other three languages. Each text pair was processed through two separate queries to the translation engine, since the two text fragments contain sufficient information to prompt an in-context translation on their own.

5 Framework

We generate SSA, LSA and ESA vectorial models for English, Romanian, Arabic, and Spanish, using the same Wikipedia 2010 versions for all the systems (e.g., the SSA, LSA and ESA relatedness measures for Spanish are all trained on the same Spanish Wikipedia version).

We construct a multilingual model by considering a word- or text-pair from a source language along

[2] For all the automatic translations we used the Google Translate service.

[3] Dataset is available for download at `lit.csci.unt.edu/index.php?P=research/downloads`

with its translations in the other languages. To evaluate this multilingual model in a way that reduces the bias that may arise from choosing one language over the other, we do the following: we start from a source language and generate all the possible combinations of this language with the available language set $\{ar, en, es, ro\}$. Within each combination, we average the monolingual model scores for the languages in this combination with respect to the target word- or text-pair 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, we aggregate the scores of the languages in that combination. 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. $\{es, ar\}$), we average the individual monolingual relatedness scores for a given word- or text-pair in this set.

Finally, to calculate the overall correlation of these generated multilingual models (one system per combination size) with the human scores, we average the correlation scores 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 we extend the multilingual model 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,

$$C_n = \{c_i \mid |c_i| = n, 0 < i < \binom{N}{n}\} \quad (1)$$

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

$$Sim_{c_i}(p) = \frac{1}{|c_i|} \sum_{l \in c_i} Sim_l(p) \quad (2)$$

where $Sim_l(p)$ is the relatedness score of the word- or 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:

$$D_i = \{\langle p, Sim_{c_i}(p)\rangle \mid p \in P\}. \quad (3)$$

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

$$Correl_{C_n}(D, G) = \frac{1}{|C_n|} \sum_{c_i \in C_n} Correl_{c_i}(D_i, G), \quad (4)$$

where $Correl$ can stand for Pearson (r), Spearman (ρ), or their harmonic mean (μ), as also reported in (Hassan and Mihalcea, 2011).

6 Evaluations

In this section we revisit the questions formulated in the introduction, and based on different experiment setups following the framework introduced in Section 5, we provide an answer to each one of them.

Does the task of semantic relatedness benefit from a multilingual representation? We evaluate the three semantic relatedness models, namely LSA, ESA and SSA on our manually constructed multilingual word relatedness ($MC30$, $WS353$) and text relatedness datasets ($LI30$), as described in Section 4.

Figure 1 plots the correlation scores achieved across all the languages against the gold standard and then averaged across all the multilingual datasets. The figure shows a clear and steady improvement (25% - 28% with respect to the monolingual baseline) achieved when more languages are incorporated into the relatedness model. It is worth noting that both the Pearson and Spearman correlations exhibit the same improvement pattern, which confirms our hypothesis that adding more languages has a positive impact on the relatedness scores. The fact that this trend is visible across all the systems supports the idea that a multilingual representation constitutes a better model for determining semantic relatedness. Furthermore, we notice that SSA is the best performing system under these settings, with a correlation improvement of approximately 15%.

To further analyze the role of the multilingual model and to explore whether some languages benefit from using this abstraction more than others, we plot the correlation scores achieved by the individual languages averaged over all the systems and the datasets in Figure 2. We notice a sharp rise in performance associated with the addition of more languages to the Arabic (42%) and the Romanian (47%) models, and a slower rise for Spanish (23%). The performance of English is also affected, but on a smaller scale (4%) when compared to the other

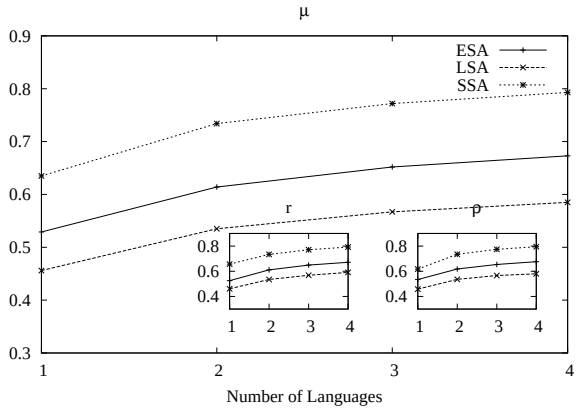

Figure 1: Manual translation - average correlation (μ, r, ρ) obtained from incorporating scores from models in other languages

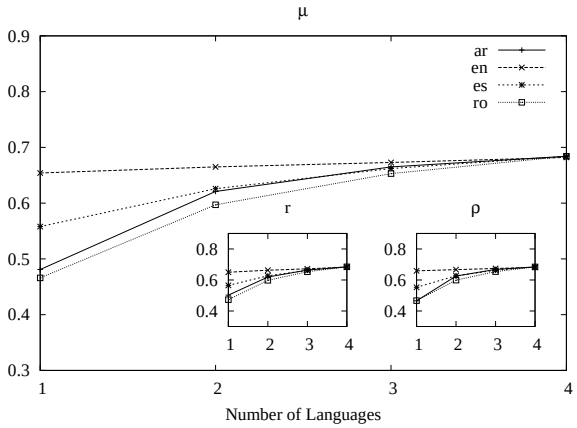

Figure 2: Manual translation - average correlation (μ, r, ρ) obtained by supplementing a source language with scores from other languages

languages. Not surprisingly, this correlates with the size of each corpus, where Arabic and Romanian are the smallest, while English is the largest.

The results support the notion that resource poor languages can benefit from languages with richer and larger resources, such as English or Spanish. Furthermore, incorporating additional languages to English also leads to small improvements, which indicates that the benefit, while disproportionate, is mutual.

Does the quality of translations affect the results? As a natural next step, we investigate the role played by the manual translations in the performance of the multilingual model. Since the previous evaluations require the availability of the word- or text-pairs in multiple languages, we attempt to see if we can eliminate this restriction 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.

We use the Google MT engine[4] to translate our multilingual datasets into the target languages (en, es, ar, and ro). We then repeat all the evaluations using the newly constructed datasets.

Figure 3 shows the correlation scores achieved across all the languages and averaged across all the multilingual datasets constructed using automatic translation. We again see a clear and steady improvement (12% - 35% with respect to the monolingual baseline) similar to the observed pattern in the corresponding manual evaluations (Figure 1). While the overall achieved performance for SSA has dropped (from $\mu = 0.793$ to $\mu = 0.71$) when compared to the manual settings, we are still able to improve over the baseline ($\mu = 0.635$). 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 miss-translations, the models overall benefit from the additional clues provided by the multilingual representation.

To explore the effect of automatic translation 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 4, in a similar fashion to Figure 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 the manual translations (Figure 2). A similar behavior is also observed in Spanish and Romanian but on a lower scale.

[4]This API is now offered as a paid service; Microsoft or Babelfish automatic translation services are publicly available.

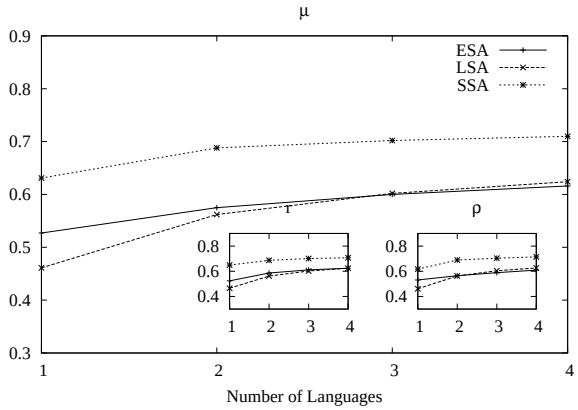

Figure 3: Automatic translation - average correlation (μ, r, ρ) obtained from incorporating scores from models in other languages

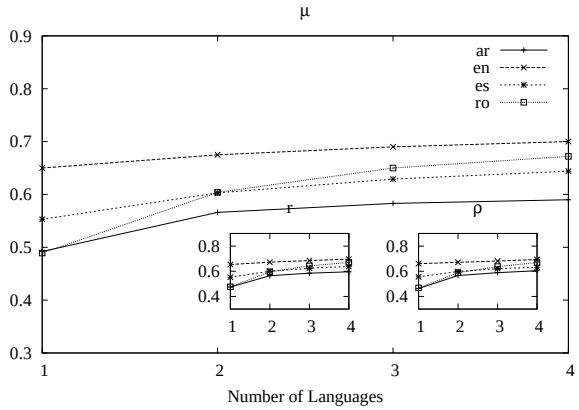

Figure 4: Automatic translation - average correlation (μ, r, ρ) obtained by supplementing a source language with scores from other languages

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 the English word in form. However, the Romanian word is highly ambiguous, and in an authoritative Romanian dictionary[5] 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 trend is positive and follows the pattern previously observed on the manually constructed datasets. This suggests that an automatic translation, even if more noisy, is beneficial and provides a way to reinforce semantic relatedness in a given language with information coming from multiple languages with no manual effort.

Do our findings hold for different relatedness datasets? At last, encouraged by the small performance difference between the use of manual versus automatic translations, we seek to explore how this multilingual model behaves under the different paradigms dictated by word relatedness versus text relatedness scenarios. Since our previous experiments were constrained to collections for which we also had a manual translation, we perform a larger scale evaluation by including automatically translated word relatedness ($RG65$) and text relatedness ($LEE50$ and $AG400$) datasets into all the languages in our language set, and repeat all the word-to-word and text-to-text evaluations.

Table 1 shows the correlation scores achieved using automatic translations on the word relatedness datasets. Most models on most datasets benefit from the multilingual representation (as shown by the figures in bold). Specifically, the SSA model has an improvement in μ of 26% for WS353 and 15% for $MC30$. 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 2 reports the results obtained for the text relatedness datasets using automatic translation. While the ESA performance suffers in the multilingual model, it is overshadowed by the improvement experienced by LSA and SSA. The multilin-

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

	r			ρ			μ		
Models	**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**

Table 1: Automatic translation - r, ρ, μ correlations on the word relatedness datasets using multilingual models.

	r			ρ			μ		
Models	**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**

Table 2: Automatic translation - r, ρ, μ correlations on the text relatedness datasets using multilingual models.

gual 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 $J\&C$ ($\mu = 0.457$).

It is important to note that the evaluation in Tables 1 and 2 are restricted to data translated from English into a target language. English, as a resource-rich language, has an extensive and robust monolingual model, yet it can still be enhanced with additional clues originating from other languages. Accordingly, we only expected small improvements in these two experiments, unlike the cases where we start from resource-poor languages such as Romanian or Arabic (see Figures 2 and 4).

7 Conclusion

In this paper, we showed how a semantic relatedness measure computed in a multilingual space is able to acquire and leverage additional information from the multilingual representation, and thus be strengthened as more languages are taken into consideration. Our experiments seem to suggest that combinations of multiple languages supply additional information to derive a semantic relatedness between texts in an automatic framework. Since establishing semantic relatedness requires us to employ cognitive processes that are in large part independent of the language that we speak, it comes at no surprise that using relatedness clues originating from more than one language allows for a better identification of relationships between texts. While efficiency may be a concern, it is worth noting that the method is highly parallelizable, as the individual relatedness measures obtained before the aggregation step can be calculated in parallel.

Notably, all the relatedness measures that we experimented with exhibited the same improvement trend. While this framework allows languages with scarce electronic resources, such as Romanian and Arabic, to obtain very large improvements in semantic relatedness as compared to the monolingual measures, improvements are also noticed for languages with richer resources such as English.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation CAREER award #0747340 and IIS award #1018613. 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

C. Banea and R. Mihalcea. 2011. Word sense disambiguation with multilingual features. In *International Conference on Semantic Computing*, Oxford, UK.

C. Banea, R. Mihalcea, and J. Wiebe. 2010. Multilingual subjectivity: Are more languages better? In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 28–36, Beijing, China, August.

R. Besançon and M. Rajman. 2002. Evaluation of a vector space similarity measure in a multilingual framework. In *Proceedings of the Third International Conference on Language Resource and Evaluation (LREC 2002)*, Las Palmas, Spain.

S. Brin, J. Davis, and H. Garcia-Molina. 1995. Copy detection mechanisms for digital documents. In *ACM International Conference on Management of Data (SIGMOD 1995)*.

A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. 1997. Syntactic clustering of the web. *Comput. Netw. ISDN Syst.*, 29(8-13):1157–1166.

A Z. Broder, P. Ciccolo, M. Fontoura, E. Gabrilovich, V. Josifovski, and L. Riedel. 2008. Search advertising using web relevance feedback. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1013–1022, New York, NY, USA. ACM.

C. Burgess, K. Livesay, and K. Lund. 1998. Explorations in context space: words, sentences, discourse. *Discourse Processes*, 25(2):211–257.

K. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

T. Cohn and M. Lapata. 2007. Machine translation by triangulation: making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735, Prague, Czech Republic.

I. Dagan, O. Glickman, and B. Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Proceedings of the PASCAL Workshop*.

D. Davidov and A. Rappoport. 2009. Enhancement of lexical concepts using cross-lingual web mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 852–861, Singapore.

L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. 2001. Placing search in context: the concept revisited. In ACM Press, editor, *The Tenth International World Wide Web Conference*, pages 406–414, Hong Kong.

E. Gabrilovich and S. Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, Hyderabad, India.

A. Goodrum. 2000. Image information retrieval: An overview of current research. *Informing Science*, 3(2):63–66.

S. Hassan and R. Mihalcea. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1201, Singapore. Association for Computational Linguistics.

S. Hassan and R. Mihalcea. 2011. Measuring semantic relatedness using salient encyclopedic concepts. *Artificial Intelligence, Special Issue*, xx(xx).

N. Heintze. 1996. Scalable document fingerprinting. In *In Proc. USENIX Workshop on Electronic Commerce*.

T. C. Hoad and J. Zobel. 2003. Methods for identifying versioned and plagiarized documents. *J. Am. Soc. Inf. Sci. Technol.*, 54(3):203–215.

A. Islam and D. Inkpen. 2006. Second order co-occurrence PMI for determining the semantic similarity of words. In *Proceedings of the Fifth Conference on Language Resources and Evaluation*, volume 2, Genoa, Italy, July.

M. Jarmasz and S. Szpakowicz. 2003. Roget's thesaurus and semantic similarity. In *Proceedings of the conference on Recent Advances in Natural Language Processing RANLP-2003*, Borovetz, Bulgaria, September.

J. Kazama, S. De Saeger, K. Kuroda, M. Murata, and K. Torisawa. 2010. A bayesian method for robust estimation of distributional similarities. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden.

T. K. Landauer, D. Laham, B. Rehder, and M. E. Schreiner. 1991. How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In *Proceedings of the 19th annual meeting of the Cognitive Science Society*, pages 412–417, Mawhwah, N. Erlbaum.

C. Leacock and M. Chodorow, 1998. *Combining local context and WordNet similarity for word sense identification*, pages 305–332.

M. D. Lee and M. Welsh. 2005. An empirical evaluation of models of text document similarity. In *Proceedings of the 27th annual meeting of the Cognitive Science Society*, pages 1254–1259, Stresa, Italy.

C. W. Leong and R. Mihalcea. 2009. Explorations in automatic image annotation using textual features. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 56–59, Suntec, Singapore, August. Association for Computational Linguistics.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries. In *Proceedings of the 5th annual international conference on Systems documentation - SIGDOC '86*, pages 24–26, Toronto, Ontario. ACM Press.

W. Li, Q. Lu, and R. Xu. 2005. Similarity based chinese synonym collocation extraction. *International Journal of Computational Linguistics and Chinese Language Processing*, 10(1).

Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138–1150, August.

D. Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, Madison, Wisconsin.

U. Manber. 1994. Finding similar files in a large file system. In *USENIX WINTER 1994 TECHNICAL CONFERENCE*, pages 1–10.

D. Metzler, Y. Bernstein, W. Bruce Croft, A. Moffat, and J. Zobel. 2005. Similarity measures for tracking information flow. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 517–524, New York, NY, USA. ACM.

D. Metzler, S. T. Dumais, and C. Meek. 2007. Similarity measures for short segments of text. In Giambattista Amati, Claudio Carpineto, and Giovanni Romano, editors, *ECIR*, volume 4425 of *Lecture Notes in Computer Science*, pages 16–27. Springer.

G. A. Miller and W. G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

G. A. Miller. 1995. WordNet: a Lexical database for english. *Communications of the Association for Computing Machinery*, 38(11):39–41.

T. Mitchell, T. Russell, P. Broomhead, and N. Aldridge. 2002. Towards robust computerised marking of free-text responses. In *roceedings of the 6th International Computer Assisted Assessment (CAA) Conference*, Loughborough, UK. Loughborough University.

M. Mohler and R. Mihalcea. 2009a. Text-to-text semantic similarity for automatic short answer grading. In *EACL*, pages 567–575. The Association for Computer Linguistics.

M. Mohler and R. Mihalcea. 2009b. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 567–575, Stroudsburg, PA, USA.

T. Pedersen, S. Patwardhan, and J. Michelizzi. 2004. WordNet::Similarity - measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04), demonstrations*, San Jose, CA.

J. Ponte and W. Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the Annual International SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281, Melbourne, Australia.

S. G. Pulman and J. Z. Sukkarieh. 2005. Automatic short answer marking. In *EdAppsNLP 05: Proceedings of the second workshop on Building Educational Applications Using NLP*, pages 9–16, Morristown, NJ, USA. Association for Computational Linguistics.

H. Rubenstein and J. B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, October.

M. Sahami and T. D. Heilman. 2006. A web-based kernel function for measuring the similarity of short text snippets. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 377–386, New York, NY, USA. ACM.

N. Shivakumar and H. Garcia-Molina. 1995. Scam: A copy detection mechanism for digital documents. In *2nd International Conference in Theory and Practice of Digital Libraries (DL 1995)*.

J. Sinclair. 2001. *Collins cobuild English dictionary for advanced learners*. Harper Collins, 3rd edition.

P. D. Turney. 2001. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning*, pages 491–502, Freiburg, Germany.

Z. Wu and M. Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133—-138, Las Cruces, New Mexico.

W. T. Yih and C. Meek. 2007. Improving similarity measures for short segments of text. In *AAAI'07: Proceedings of the 22nd national conference on Artificial intelligence*, pages 1489–1494. AAAI Press.

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*.