Word semantic similarity for morphologically rich languages

Word Semantic Similarity for Morphologically Rich Languages

Kalliopi Zervanou[1], **Elias Iosif**[2] **and Alexandros Potamianos**[3]
[1] SAIL, University of Southern California, CA 90089, USA
[2] Athena Research and Innovation Center, Maroussi 15125, Athens, Greece
[3] School of ECE, National Technical Univ. of Athens, Zografou 15780, Greece

Abstract

In this work, we investigate the role of morphology on the performance of semantic similarity for morphologically rich languages, such as German and Greek. The challenge in processing languages with richer morphology than English, lies in reducing estimation error while addressing the semantic distortion introduced by a stemmer or a lemmatiser. For this purpose, we propose a methodology for selective stemming, based on a semantic distortion metric. The proposed algorithm is tested on the task of similarity estimation between words using two types of corpus-based similarity metrics: co-occurrence-based and context-based. The performance on morphologically rich languages is boosted by stemming with the context-based metric, unlike English, where the best results are obtained by the co-occurrence-based metric. A key finding is that the estimation error reduction is different when a word is used as a feature, rather than when it is used as a target word.

Keywords: distributional semantic models, lexical semantics, morphology, morphologically rich languages

1. Introduction

Semantic similarity is the building block for numerous applications of natural language processing, such as grammar induction (Meng and Siu, 2002) and affective text categorisation (Malandrakis et al., 2011). Recently, there has been an increased research interest in devising data-driven approaches for estimating semantic similarity between words, phrases, and sentences. However, data-driven approaches to semantic similarity estimation mainly focus on the English language (e.g. the SemEval sentence-level semantic similarity challenges). Little evidence currently exists on how these algorithms port to other languages, especially languages characterised by greater syntactic and/or morphological variability than English.

Distributional semantic models (DSMs) (Baroni and Lenci, 2010) are based on the distributional hypothesis of meaning (Harris, 1954) in assuming that semantic similarity between words is a function of the overlap of their linguistic contexts. DSMs can be categorised into *structured models*, which employ syntactic relationships between words (Grefenstette, 1994; Baroni and Lenci, 2010), and *unstructured models*, which employ a bag-of-words model (Iosif and Potamianos, 2010). DSMs are typically constructed from co-occurrence statistics of word tuples, extracted either of existing corpora, or corpora specifically harvested from the web. In a recent work, Iosif and Potamianos (2013) proposed general-purpose, language-agnostic algorithms for estimating semantic similarity, without any linguistic resources other than a corpus created via web queries. They demonstrate that the main reason behind the success of the proposed methods lies in the construction of semantic networks and semantic neighbourhoods that capture smooth co-occurrence and context similarity statistics.

In this work, we investigate the performance of these network DSMs on European languages with richer morphology than English. Our objective is not only to report the performance of the semantic similarity algorithms on these languages, but also to investigate the role of morphology on performance. Moreover, we propose a selective stemming algorithm, so as to reduce the vocabulary size in our corpus, while the overall semantic distance between the original and the stemmed version of the text is controlleded by a semantic distortion metric.

In addition to English, the languages investigated in this work are German and (Modern) Greek. Although all three languages are synthetic/fusional languages (Sapir, 1921), German (Bane, 2008) and Greek present a richer morphology than English, with Greek being the richest of the three in terms of inflectional variation (Koliopoulou, 2013).

2. Morphologically Rich Languages

Morphologically rich languages are characterised by highly productive morphological processes (inflection, agglutination, compounding) which may result in a very large number of word forms for a given root form (Sarikaya et al., 2009; Tsarfaty et al., 2013). Moreover, the lexical information for each word form in a morphologically rich language is augmented with other types of information, such as those related to its grammatical role and syntactic function or temporal information and pronominal clitics.

Using the communication protocol analogy for language, one may assume that on average all languages have the same *semantic entropy*, namely semantic information is

transmitted with approximately the same efficiency in all languages. Thus, a reduction in the number of symbols available in one layer of the language code (e.g., phonology) will be typically compensated in another layer (e.g., morphology). For example, English and German are richer phonemically than Greek, but poorer in terms of inflectional affixes and while syntax in Greek has a relatively flexible constituent ordering, German and English have very strict syntactic constraints regarding constituent order.

Although the overall efficiency of language as a communication protocol may remain intact, morphologically rich languages pose challenges to lexical semantics and natural language processing (NLP) algorithms. This is due to the fact that morphological productivity in those languages results in an increase of vocabulary size for NLP applications. From a lexical semantics perspective, the reduction of symbols at a given lexical or syntactic layer of the language and the increase in ambiguity that this reduction entails at the given layer, has an adverse impact on the efficiency of semantic similarity estimations. In particular, the main features used for estimating the semantic similarity between two words are:

- their direct co-occurrence in text, and
- the overlap of their context

Although these features remain rather unaffected by lexical variation and ambiguity, the accuracy in statistical similarity estimations is gravely impacted. For example, when the number of words increases, co-occurring words are also more diverse and similarity estimations should rely on sparser examples. In addition to the number of words, the second feature, the context is also affected by syntactic variation, since context can be more or less variable depending on the language. For these reasons, we need to devise semantic similarity algorithms that reduce vocabulary size with minimum increase in ambiguity and respective distortion in the semantics of the corpus.

3. Corpus-Based Similarity Metrics

In this section, we discuss in more detail the methods used in building corpus-based semantic similarity models. First, we briefly present the motivation and the definition of co-occurence and context-based similarity metrics and, subsequently, we present the rationale and the methodology in building network-based DSMs and how these have been implemented in this work.

3.1. Co-occurrence-based metrics

The underlying assumption of co-occurrence-based metrics is that the co-appearance of words in a specific contextual environment indicates semantic relatedness. In this work, we employ a widely used co-occurrence-based metric, the Dice coefficient metric. The Dice coefficient between words w_i and w_j is defined as follows:

$$D(w_i, w_j) = \frac{2f(w_i,w_j)}{f(w_i)+f(w_j)}, \quad (1)$$

where $f(.)$ denotes the frequency of word occurrence/co-occurrence. Here, the word co-occurrence is considered at the sentential level, while D can be also defined with respect to broader contextual environments, such as the paragraph level (Véronis, 2004).

3.2. Context-based metrics

The fundamental assumption behind context-based metrics is that *similarity of context implies similarity of meaning* (Harris, 1954). In this work, a contextual window of size $2H+1$ words is built for the word of interest w_i and lexical features are extracted. For every instance of w_i in the corpus, H words left and right of w_i compose a feature vector v_i. For a given value of H, the context-based semantic similarity between two words, w_i and w_j, is computed as the cosine of their feature vectors:

$$Q^H(w_i, w_j) = \frac{v_i.v_j}{||v_i||\ ||v_j||}. \quad (2)$$

The elements of feature vectors can be weighted according to various schemes (Iosif and Potamianos, 2010). In the work presented in this paper, we use a binary scheme.

3.3. Network-based DSMs

In this section, we summarise the main ideas of network-based DSMs, as proposed in Iosif and Potamianos (2013). The rationale behind the implementation of network-based DSMs has been twofold. First, the network is a parsimonious representation of the corpus statistics pertaining to semantic similarity estimation of word-pairs. Second and most importantly, the network representation allows for discovering relations that are not directly observable in the data; such relations emerge via the systematic covariation of similarity metrics.

The proposed network is defined, under a symmetric similarity metric, as an undirected graph $F = (V, E)$, where the set of vertices V consists of all words in our vocabulary set L, and the set of edges E consists of all the links between these vertices. The edges between words in the network are determined and weighted according to the pairwise semantic similarity of the vertices.

In building this network, for each reference word w_i that is included in the vocabulary set L, $w_i \in L$, we consider a sub-graph of F, $F_i = (N_i, E_i)$, where the set of vertices N_i includes n members of L in total, which, in turn, are linked to w_i via an E_i set of edges. This F_i sub-graph is referred to as the *semantic neighbourhood* of w_i. The members of N_i (i.e. the neighbours of w_i) are selected

according to a semantic similarity metric with respect to w_i, either co-occurrence-based D, or context-based Q^H, as defined in Eq. (1) and Eq. (2), above. Thus, the n most similar words to w_i are selected.

The notion of *semantic neighbourhood* (Iosif and Potamianos, 2013) is utilised in this work in two semantic similarity metrics, the estimation of *maximum similarity of neighbourhoods* and the *sum of squared neighbourhood similarities* metric. In the following subsections, we present these two semantic neighbourhood metrics in more detail.

3.3.1. Maximum Similarity of Neighbourhoods

This metric is based on the hypothesis that the similarity of two words, w_i and w_j, can be estimated by *the maximum similarity of their respective sets of neighbours*, and is defined as follows:

$$M_n(w_i, w_j) = \max\{\alpha_{ij}, \alpha_{ji}\} \tag{3}$$

where:

$$\alpha_{ij} = \max_{x \in N_j} S(w_i, x), \quad \alpha_{ji} = \max_{y \in N_i} S(w_j, y)$$

α_{ij} (or α_{ji}) denotes the maximum similarity between w_i (or w_j) and the neighbours of w_j (or w_i). This similarity is computed according to a similarity metric S, in this work either D, or Q^H. N_i and N_j denote the sets of neighbours for w_i and w_j, respectively.

The definition of maximum similarity M_n is motivated by the maximum sense similarity assumption (Resnik, 1995). In our work, the underlying assumption is that the most salient information in the neighbours of a word constitute semantic features denoting the senses of this word.

3.3.2. Sum of Squared Neighbourhood Similarities

While in *maximum similarity M_n metric* we consider only the most salient information in the neighbours of a word, i.e. only the most similar neighbours in the N_j set to w_i and the most similar neighbours in the N_i set to w_j, in the *sum of squared neighbourhood similarities* metric, we exploit all information available in the N_i and N_j neighbourhood sets of w_i and w_j for the estimation of their semantic similarity.

In particular, the *sum of squared neighbourhood similarities* estimation, for a given pair of words w_i and w_j, is defined as follows:

$$E_n^\theta(w_i, w_j) = \left(\sum_{x \in N_j} S(w_i, x)^\theta + \sum_{y \in N_i} S(w_j, y)^\theta \right)^{\frac{1}{\theta}} \tag{4}$$

where N_i is the set of neighbours of word w_i, and S, in this work, is the Dice similarity metric, defined in Eq. (1), above. The contribution of all neighbours to the final similarity score $E_n^\theta(w_i, w_j)$ is performed by summing the squares ($\theta = 2$) of similarities between w_i and the neighbours of w_j. The same calculation is repeated for w_j and the neighbours of w_i, so as to make $E_n^\theta(w_i, w_j)$ symmetric. The $E_n^{\theta=2}$ metric is unbounded, because the yielding similarity scores range within $[0, \infty]$. The contribution of each word–to–neighbour similarity is non-linearly weighted, using the square of the respective similarity score. The motivation behind using $\theta > 1$ is that more similar words in the neighbourhoods should be weighted more in the final similarity decision[1]. Note that as θ goes to ∞, E_n^θ and M_n become equivalent.

4. DSMs for Morphologically Rich Languages

Typically, NLP and information retrieval applications address the issue of morphological word form variation by applying text normalisation techniques, such as stemming or lemmatisation. Information retrieval research, for example, has since its very early beginnings introduced stemming as a preprocessing step in building document and query representations. Currently, both stemming and lemmatisation constitute a text preprocessing step in many NLP applications, so as to reduce the total number of surface word forms, both in cases where word forms are used as targets, such as in statistical language modelling, as well as in cases where word forms are used as features, such as in semantic similarity estimation using context-based metrics. The main challenge in developing DSMs for morphologically rich languages lies in the large vocabulary size which affects the efficiency of semantic similarity estimations, since these typically model each word form as a separate word.

A reduction of vocabulary size by means of stemming may succeed in improving statistical estimations of word occurrence or co-occurrence. However, reducing words to a stemmed or (less-so) lemmatised form also introduces a shift in semantics, where for example *people* and *peoples* have quite different meaning. We therefore observe two opposing forces in action, one where stemming has a positive effect in reducing *estimation error* by reducing the number of word forms and respective sparsity, and one where stemming introduces *semantic distortion* by grouping morphologically similar words into potentially semantically disparate groups.

In this section, we discuss our proposal for explicitly optimising the decision on which words to stem using a semantic distortion metric. Suppose that we have a set L of words, where $w_i \in L$ and a set K of stemming rules,

[1] Despite the resemblance between the $E_n^{\theta=2}$ metric and the Euclidean distance, no assumption is adopted here about the semantic neighbourhoods being metric spaces under S.

where $r_k \in K$. Then the stemmed version of a word w_i would be defined as $l_k(w_i)$, where $l_k(.)$ is the stem of w_i after rule r_k is applied.

We also have a corpus-based semantic similarity metric $S(w_i, w_j)$ between two words – or word stems, w_i and w_j, normalised in $[0, 1]$. For each rule r_k, the average semantic distortion C is computed as the average distance between the original word w_i and its stemmed form $l_k(w_i)$. Specifically:

$$C(r_k) \propto \frac{\sum_{i=1}^{N} f(w_i)[1 - S(w_i, l_k(w_i))]}{\sum_{i=1}^{N} f(w_i)[1 - \delta(w_i, l_k(w_i))]} \quad (5)$$

where $f(w_i)$ assigns a weight based on the frequency of the word form w_i in the corpus, and $\delta(.,.)$ is Kronecher's delta, a value that is set to 1, when $w_i \equiv l_k(w_i)$, and 0, otherwise. Estimating the semantic similarity between a word w_i and its stemmed form $l_k(w_i)$ is not straightforward, because the stem $l_k(w_i)$ typically does not appear in the corpus. Instead, we define $S(w_i, l_k(w_i))$ as the average similarity between w_i and the subset W of all words in the vocabulary L sharing the same stem after r_k is applied to w_i.

The semantic distortion $C(r_k)$ is computed for each stemming rule r_k. Then, the rules are ranked according to their respective $C(r_k)$, in increasing order.

5. Experimental Procedure and Results

5.1. Baseline Experiments

For our baseline experiments, we have used a vocabulary set of $10,000$ noun word forms in all three languages. For each word form in this set, an individual query was formulated and the 1000 top ranked results, i.e. document snippets were retrieved using the Yahoo! search engine[2]. The respective corpus for each language is created by aggregating the web snippets for all nouns in our vocabulary set.

Subsequently, we estimated semantic similarity, with and without stemming, using Dice coefficient for co-occurrence similarity (D), cosine similarity with window size 1 ($Q^{H=1}$) for context-based similarity, and the two variants of network-based metrics, i.e. M_n and $E_n^{\theta=2}$. For stemming, we have used for English and German, the respective Snowball stemmers[3]. For Greek there was no general purpose stemmer available, thus no results for Greek are reported in this section.

The performance of these similarity metrics and the effect of stemming were evaluated for the task of semantic similarity between nouns. Pearson's correlation coefficient was used as evaluation metric to compare estimated similarities against human ratings. For English and German, we used the evaluation datasets presented in Rubenstein and Goodenough (1965) and Gurevych (2005), respectively. Of these evaluation datasets, we included only those word pairs that were covered by our networks: 57 out of 65 pairs for English, and 63 out of 65 pairs for German.

Language	Similarity metric	Corpus processing	
		None	Stem.
English	D	0.60	0.52
	$Q^{H=1}$	0.55	0.02
	M_n	**0.87**	0.61
	$E_n^{\theta=2}$	0.86	0.57
German	D	0.24	0.27
	$Q^{H=1}$	0.41	0.21
	M_n	0.65	**0.68**
	$E_n^{\theta=2}$	0.67	0.65

Table 1: Pearson's correlation to human similarity judgements for unstemmed and stemmed data, and a vocabulary of 10K words, nouns only.

The results in semantic similarity estimation are shown on Table 1. For the network-based metrics, M_n and $E_n^{\theta=2}$, the performance is reported for neighbourhood size $n = 100$. For neighbour selection both methods used Dice coefficient D, while for similarity estimation, we have used cosine similarity $Q^{H=1}$ for M_n and Dice coefficient for $E_n^{\theta=2}$. The effect of the neighbourhood size n and the combinations of D and $Q^{H=1}$ in neighbour selection and similarity computation are discussed in detail in Iosif and Potamianos (2013).

For the co-occurrence-based metric D, stemming appears to slightly improve the performance for German. It is also quite notable that when stemming is used for English the performance is consistently worse in all four metrics, with the context-based metric $Q^{H=1}$ presenting the worst results. The performance of the context-based metric $Q^{H=1}$ for German deteriorates by stemming, albeit less than in English.

Both network-based metrics, M_n and $E_n^{\theta=2}$ perform better than D and $Q^{H=1}$. However, use of stemming degrades the performance for English in both M_n and $E_n^{\theta=2}$. For German, the use of stemming with M_n improves performance, whereas stemming with $E_n^{\theta=2}$ slightly decreases performance. Although not reported here, the effect of stemming is rather beneficial for Greek, as discussed in the next section.

Overall, the stemming effect seems to be a function of the used similarity metric and the morphological richness of the language under investigation.

[2] http://www.yahoo.com//

[3] snowball.tartarus.org

5.2. Stemming Rules Ranking Experiments

In order to acquire a better picture of the effects of stemming and the resulting semantic distortion, we attempted to cover the largest possible set of words in our vocabulary and of all grammatical parts of speech, rather than nouns only. For this purpose, we compiled our vocabularies for each of the languages under investigation, using the intersection of the vocabulary of the entire set of Wikipedia pages and the aspell dictionaries. This process resulted in a vocabulary of 135,436 word forms for English, 332,456 word forms for German and 407,752 word forms for Greek. Our corpus is acquired using the same methodology as in the baseline experiments, namely by aggregating the 1000 top ranked web snippets retrieved for each word form in our vocabulary set. For stemming, we have used for English and German, the Snowball stemmers, whereas for Greek, we have developed a general purpose stemmer[4].
Similarity metrics are used in two stages:

(i) First, in order to compute the individual rule semantic distortion, we need to estimate the similarity between a word and its stem, as defined in Eq.(5). For estimating the similarity between the surface word form w_i and its stem $l_k(w_i)$, $S(w_i, l_k(w_i))$, we use $E_n^{\theta=2}$, because, as also shown on Table 1, network-based metrics achieve better performance.

(ii) Second, in the final similarity metric estimation between a pair of words, as defined in Eq.(1-4), co-occurrence-based metric Dice coefficient D and context-based $Q^{H=1}$ similarity metrics are used.

For the context-based metric $Q^{H=1}$, we have experimented with three stemming scenarios:

(i) using stemming only for the target words (i.e. the pairs)

(ii) using stemming only for the context words

(iii) stemming both the target and the context words.

Once the semantic distortion is estimated, the stemming rules are ranked by increasing distortion value. Subsequently, a percentage of stemming rules is selected by thresholding the semantic distortion criterion C.
Results are reported for co-occurrence D and context-based $Q^{H=1}$ similarity metrics. The evaluation datasets used in these experiments were the same as those used in our baseline experiments. In addition, due to lack of evaluation datasets for the Greek language, a new dataset of 200 pairs was compiled by experts and results are reported also on these datasets. The semantic similarity performance reported is Pearson's correlation with human similarity judgement.

[4]This stemmer is an extended and modified version of the Greek Stemmer developed by Ntais (2006).

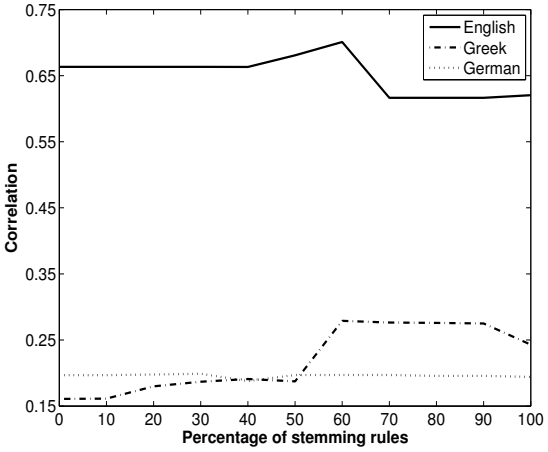

Figure 1: Performance of co-occurrence metric D as a function of the percentage of stemming rules applied.

The obtained correlation for the co-occurrence-based metric is depicted in Fig. 1, as a function of the percentage of the applied stemming rules. This is shown for all three languages. We observe that the achieved correlation is significantly higher for English, compared to Greek and German. Moreover, it is shown that at 60% of the rules, the performance is slightly improved for both English and Greek, unlike German, where stemming does not seem to affect an overall low performance.
Fig. 2 illustrates the obtained correlation, as a function of the percentage of the applied stemming rules, for the context-based metric $Q^{H=1}$. Fig. 2(a) refers to the application of stemming to the target words only, Fig. 2(b), to context words only, and Fig. 2(c), to both the target and the context words. We observe that Greek performs overall better in all three approaches, while German achieves the lowest performance.
The best results are achieved by the application of stemming to the target words only (cf. Fig. 2(a)), where a slight gain in performance is achieved for Greek, when 30% of the top stemming rules is applied. Rule selection seems to play a more significant role in the stemming of contextual features approach, as this is illustrated in Fig. 2(b), where the performance steeply degrades, especially for English, above the 40% of stemming rules.

6. Conclusions

In this work, we have investigated the role of morphology in the performance of unstructured DSMs. For this purpose, in addition to English, we have experimented with two European languages, German and Greek, which are

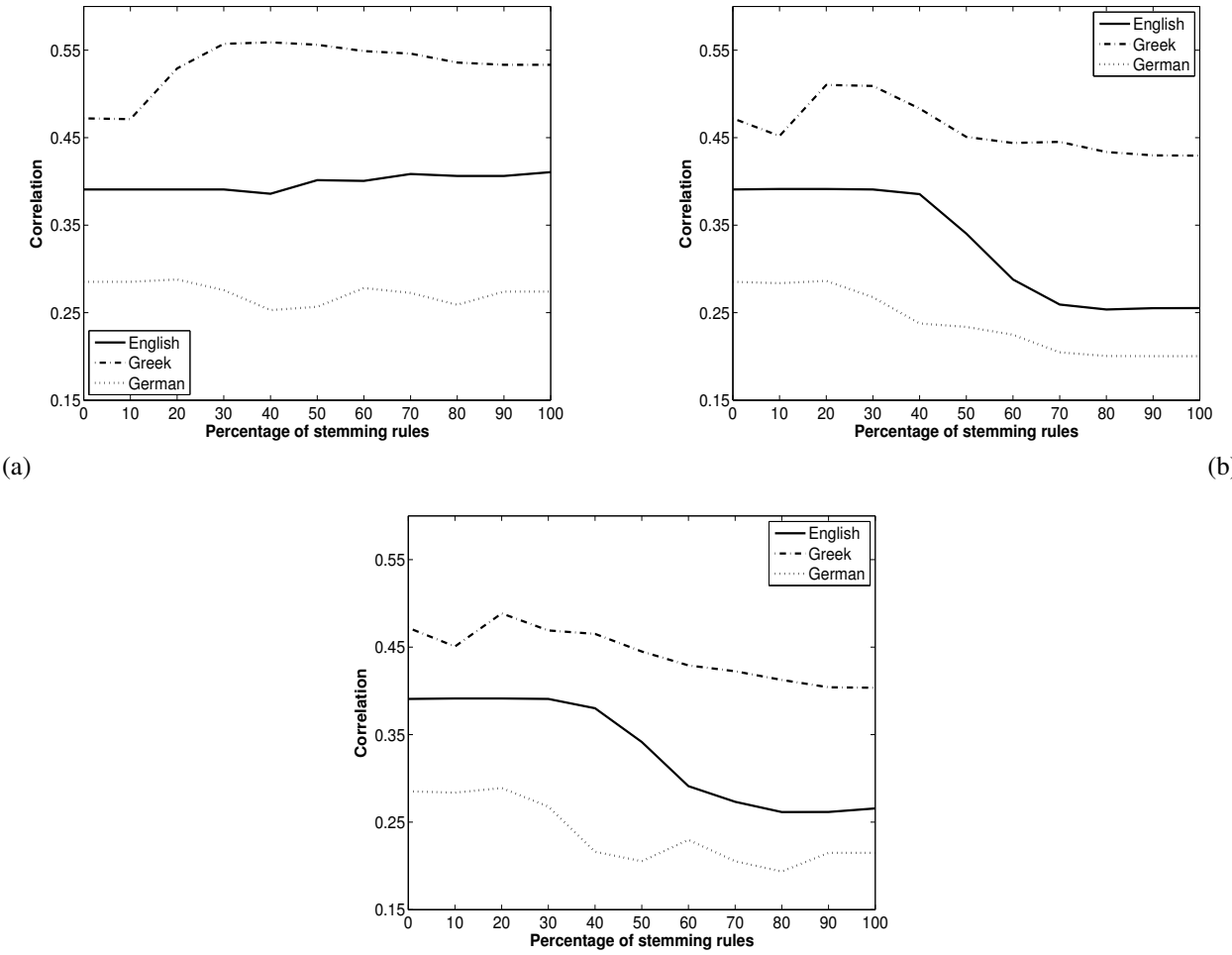

Figure 2: Performance of context-based metric $Q^{H=1}$ as a function of the percentage of stemming rules applied. Using stemming: (a) only for the target words, (b) only for context words, (c) for both the target and the context words.

characterised by richer morphology than English. We proposed a methodology for selective stemming, based on a semantic distortion metric which counterbalances the similarity estimation error reduction, achieved by text normalisation techniques, such as stemming, with the semantic distortion that such a process entails.

We observed that selective stemming application with the co-occurrence-based metric is beneficial for languages characterised by a poorer morphology, such as English, while stemming in combination with a context-based metric deteriorates performance. This phenomenon is reversed in a morphologically richer language, such as Greek, where poor performance with the co-occurrence-based metric is greatly improved with the context-based metric. This may be attributed to the number of features considered for similarity estimation. In particular, in the co-occurrence-based metric, the estimation of similarity relies solely on the statistics of the co-occurring words. Conversely, in the context-based metric, the number of lexical features considered is much larger. Thus, selective stemming for a morphologically rich language reduces the vocabulary size and improves the similarity estimation. The performance of German is low overall and stemming does not seem to affect much the performance. A possible explanation for this might lie in morphological phenomena, such as compounding, which also increase vocabulary size, but cannot be addressed by a simple text normalisation technique, such as stemming. The formation of ad-hoc single word compounds is a particular characteristic of German that is potentially responsible for the low performance in similarity estimations.

The key observation is that the estimation error reduction is different, when a word is used as a feature, where errors are averaged typically among many different fea-

tures, than when it is used as a target word. For example, in the context-based semantic similarity estimation it makes sense to perform stemming for the words proper that we are estimating the semantic similarity for, but less so for the context feature vector, where we have potentially enough statistics, especially if a large window is used.

As future work, we plan to investigate more refined stemming rule selection and semantic similarity performance for the network-based metrics when using stemming. Additionally, we would like to investigate in more detail the effect of stemming in relation to the particularities of derivational vs. inflectional morphology, and the role of prefixes in semantics. Finally, we are interested in investigating other languages, such as Romance languages, or languages characterised by agglutinative morphological phenomena, such as Turkish.

7. Acknowledgements

Elias Iosif and Alexandros Potamianos were partially funded by the EU-IST FP7 PortDial project ("Language Resources for Portable Multilingual Spoken Dialog Systems"), grant number 296170. The authors wish to thank Maria Giannoudaki for the development of the Greek evaluation dataset and preliminary experiments, and Ioannis Klasinas for corpora creation.

8. References

Bane, M. (2008). Quantifying and measuring morphological complexity. In *Proc. of the 26th West Coast Conference on Formal Linguistics*, pages 67–76.

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers.

Gurevych, I. (2005). Using the structure of a conceptual network in computing semantic relatedness. In *Proc. of the 2nd International Joint Conference on Natural Language Processing*.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Iosif, E. and Potamianos, A. (2010). Unsupervised semantic similarity computation between terms using web documents. *IEEE Transactions on Knowledge and Data Engineering*, 22(11):1637–1647.

Iosif, E. and Potamianos, A. (2013). Similarity computation using semantic networks created from web-harvested data. *Natural Language Engineering* (DOI: 10.1017/S1351324913000144).

Koliopoulou, M. (2013). *Issues of Modern Greek and German compounding: a contrastive approach*. Ph.D. thesis, University of Patras.

Malandrakis, N., Potamianos, A., Iosif, E., and Narayanan, S. (2011). Kernel models for affective lexicon creation. In *Proc. Interspeech*, pages 2977–2980.

Meng, H. and Siu, K.-C. (2002). Semi-automatic acquisition of semantic structures for understanding domain-specific natural language queries. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):172–181.

Ntais, G. (2006). Development of a stemmer for the Greek language. Master's thesis, Stockholm University – Royal Institute of Technology.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of International Joint Conference for Artificial Intelligence*, pages 448–453.

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

Sapir, E. (1921). *Language: An Introduction to the Study of Speech*. New York: Harcourt, Brace, 1921; Bartleby.com, 2000.

Sarikaya, R., Kirchhoff, K., Schultz, T., and Hakkani-Tur, D. (2009). Introduction to the special issue on processing morphologically rich languages. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5):861–862.

Tsarfaty, R., Seddah, D., Kuebler, S., and Nivre, J. (2013). Parsing Morphologically Rich Languages: Introduction to the Special Issue. *Computational Linguistics*, 39(1):15–22.

Véronis, J. (2004). Hyperlex: Lexical cartography for information retrieval. *Computer Speech and Language*, 18(3):223–252.