Combining dependency information and generalization in a pattern-based approach to the classification of lexical-semantic relation instances

Silvia Necşulescu*, **Sara Mendes***†, **Núria Bel***
*Universitat Pompeu Fabra
Roc Boronat, 138, Barcelona, Spain
{silvia.necsulescu, sara.mendes, nuria.bel}@upf.edu

† Centro de Linguística da Universidade de Lisboa
Avenida Professor Gama Pinto, 2, Lisboa, Portugal
sara.mendes@clul.ul.pt

Abstract

This work addresses the classification of word pairs as instances of lexical-semantic relations. The classification is approached by leveraging patterns of co-occurrence contexts from corpus data. The significance of using dependency information, of augmenting the set of dependency paths provided to the system, and of generalizing patterns using part-of-speech information for the classification of lexical-semantic relation instances is analyzed. Results show that dependency information is decisive to achieve better results both in precision and recall, while generalizing features based on dependency information by replacing lexical forms with their part-of-speech increases the coverage of classification systems. Our experiments also make apparent that approaches based on the context where word pairs co-occur are upper-bound-limited by the times these appear in the same sentence. Therefore strategies to use information across sentence boundaries are necessary.

Keywords: lexical-semantic relations, pattern-based classification systems, information extraction

1. Introduction

The automatic classification of relation instances is an important area for many NLP applications and tasks such as the automatic construction of relational language resources like WordNet (Miller, 1995) or for inference systems in the area of information retrieval and question answering (Pasca and Harabagiu, 2001).

The present article addresses the classification of word pairs as instances of lexical-semantic relations. For instance, the word pairs *(table:leg)* and *(pan:handle)* are both instances of the relation of meronymy: since *tables* have *legs* and *pans* have *handles*; *leg* and *handle* are meronyms of *table* and *pan*, respectively.

Hearst (1992)'s seminal work in this area opened a line of research followed by many authors who have focused on the identification of specific relations like *hypernymy* (Hearst, 1992; Ravichandran and Hovy, 2002; Snow et al., 2004), *meronymy* (Berland and Charniak, 1999; Girju et al., 2006), *cause* (Girju, 2003) and *antonymy* (Lin et al., 2003), resulting in dedicated approaches to each type of semantic relation. Considering the vast range of semantic relations connecting words, creating a different system for the identification of each type of semantic relation available will tend to result in approaches lacking in generalization, besides being very inefficient as a global approach. Turney (2008b) proposes a unified treatment for all semantic relations, approaching the classification of relation instances as an analogy task: a word pair has the relation r if it is similar to a known instance of relation r (see Section 2 for a more detailed presentation of the general lines of this approach). A similar method, known as distant supervision, was applied by Mintz et al. (2009) for classifying relation instances regarding *world knowledge*. These are relations holding between named-entities, and typically used to populate knowledge databases such as Freebase (Bollacker et al., 2008) and Yago (Suchanek et al., 2007).

Although the identification of these two types of relations, lexical-semantic relations and world-knowledge relations, has been treated in a similar way in the literature, by relying on a set of example word pairs and their context of co-occurrence as the source of information provided to automatic systems, each of these groups has different properties, which are reflected in the distributional behavior of instances of these relations and thus in corpora data.

As aforementioned, world-knowledge relations typically hold between two named-entities, while lexical-semantic relations link words in any parto-of-speech. Moreover, a pair of named-entities can instantiate more than one semantic relation (Hoffmann et al., 2011; Surdeanu et al., 2012), which is never the case for lexical-semantic relations. For instance the entities *Obama* and *United States* instantiate the relation *BornIn*, as well as the relation *EmployedBy*. Moreover, world-knowledge relations tend to be explicitly expressed in language data in sentences like *Obama was born in the United States*, for instance, in which the explicit context *X was born in Y* expresses the relation *BornIn* holding between the two entities *Obama* and *United States*. This is generally not the case for lexical-semantic relations, as these relations tend not to be explicitly expressed in language data, although they can be inferred from the set of contexts in which their instances occur.

For instance, Girju et al. (2003) extracted patterns of co-occurrence contexts for the relation of meronymy. 92.15% of these patterns were phrase-level patterns, such as *X of Y* where X is the part and Y is the whole. These phrase-based patterns often match contexts expressing more than one relation type which makes lexical-semantic relation instances harder to classify.

In this paper we focus on lexical-semantic relations and we apply an approach similar to Turney (2008b)'s and Mintz et al. (2009)'s for the classification of instances of these semantic relations: starting from a set of examples of a target relation r, a supervised system automatically learns which patterns of co-occurrence contexts, as collected from corpus data, express the relation r; new word pairs are classified based on the context where they co-occur in corpus. To accomplish this, the contexts where two words co-occur are transformed in a pattern by replacing each target word with a slot. These patterns are then used as cues to gather distributional information from corpus, and provide it to the classification system.

Naturally, the results of systems based on this approach depend on the amount of contexts where each word pair co-occurs and, due to the ubiquitous phenomenon of data sparseness, they are characterized by low recall scores. In the work presented in this paper, we focus on strategies for improving recall scores of pattern-based classification systems using only information extracted from corpora.

We use dependency paths, i.e. sequences of dependency relations, connecting two target words in a dependency graph output by a dependency parser as the context of co-occurrence of word pairs. Current systems based on this type of context use only the shortest dependency paths connecting two target words (Snow et al., 2004; Mintz et al., 2009; Wu and Weld, 2010). One of the strategies proposed here is to leverage all the dependency paths up to three edges instead of only the shortest ones to create patterns of co-occurrence contexts. Additionally, to find similarities between slightly varying contexts of co-occurrence, and consequently increase the coverage of the automatically acquired patterns, we also test a generalization strategy based on part-of-speech information for creating more general patterns.

Therefore to overcome this upper-limit, strategies to use information across sentence boundaries are necessary.

The rest of the paper is structured as follows: Section 2. introduces previous work addressing the general research lines followed in this paper to address the general scientific problem described in Section 3.; Section 4. describes a system combining two strategies to improve recall scores of pattern-based systems for the classification of relation instances; Section 5. presents the experimental setup whose results are presented in Section 6. and discussed in Section 7.; final remarks are presented in Section 8.

2. Previous work

Hearst (1992)'s work pioneered the automatic extraction of instances of semantic relations. This author manually created a set of lexical-syntactic patterns to find word pairs instantiating a relation of hypernymy. Berland and Charniak (1999) developed a similar approach for identifying meronyms. These approaches are characterized by good precision but a very low recall due to the large variation of contexts expressing a semantic relation. In fact, although, for each relation there may be a small set of very precise patterns, such as *X is a part of Y* for the relation of *meronymy*, these do not necessarily occur with all the instances of the relation in a corpus. This way, to increase the recall of this type of approach additional patterns have to be developed, which increases development costs exponentially, as these patterns tend to be less frequent, thus having to be more numerous to effectively affect the coverage of the system.

To overcome these limitations and the high cost of manual development, approaches for automatically extracting patterns of co-occurrence have been developed. Snow et al. (2004) used WordNet to automatically acquire patterns based on dependency relations for hypernym extraction, Davidov and Rappoport (2006) used symmetry patterns and high frequency words for co-hyponym extraction, while Girju et al. (2006) used general patterns - patterns having high coverage of the corpus, but low precision -, manually annotated with semantic information extracted from WordNet for meronym detection.

However, each of these authors focuses on a single specific relation, which results in the definition of different types of patterns for automatically identifying instances of each semantic relation. To avoid this, Turney (2008b) proposed a uniform approach for the classification of semantic relation instances. According to this author, the automatic identification of instances of any semantic relations is subsumed as an analogy task: given a target relation r, a word pair (x, y) can be labeled as an instance of the semantic relation r if (x, y) is analogous to a word pair $(x, y)_r$, instance of r. This approach builds on the *distributional hypothesis* which states that when two words have similar distributions they tend to share aspects of meaning. The *latent relation hypothesis* (Turney, 2005) reformulates the distributional hypothesis for pairs of words: pairs of words that co-occur in similar contexts tend to have the same lexical-semantic relation. Therefore, to recognize analogous word pairs, the similarity between their distributional behavior is calculated based on patterns extracted from the contexts in which both members of the pair co-occur. Starting from a set of examples, this author automatically acquired contexts of co-occurrence from an input corpus and used them to generate lexicalized patterns of co-occurrence.

Presently, there are two mainstream lines of research regarding the acquisition of patterns of co-occurrence. Turney's work is based on surface patterns acquired from a very large corpus ($\sim$50 Gb of text) (Turney, 2005; Turney, 2006b; Turney, 2008a; Turney, 2008b). Other works use dependency information gathered from a parsed corpus (Snow et al., 2004; Mintz et al., 2009; Wu and Weld, 2010) to reduce the requirements regarding the dimensions of the input corpus. In this kind of approach, the patterns of co-occurrence are extracted from dependency paths relating two entities in a dependency graph of a sentence in which they co-occur. The general approach based on this type of information relies only on the shortest paths between each pair of entities. To increase the precision of the system, neighboring "window" tokens, i.e. nodes that are not within the dependency path connecting two words, but are connected to one of the nodes in the dependency graph, are included into the patterns of co-occurrence.

Both types of patterns used in previous work, surface pat-

$screwdriver \xrightarrow{\text{conj}} X_N \xleftarrow{\text{obj}} handle_V \xrightarrow{\text{obj}} Y_N \xleftarrow{\text{mod}} other_J$
$screwdriver \xrightarrow{\text{conj}} X_N \xleftarrow{\text{obj}} V \xrightarrow{\text{obj}} Y_N \xleftarrow{\text{mod}} other_J$
$screwdriver \xrightarrow{\text{conj}} X_N \xleftarrow{\text{obj}} handle_V \xrightarrow{\text{obj}} Y_N$
$screwdriver \xrightarrow{\text{conj}} X_N \xleftarrow{\text{obj}} V \xrightarrow{\text{obj}} Y_N$
$X_N \xleftarrow{\text{obj}} handle_V \xrightarrow{\text{obj}} Y_N \xleftarrow{\text{mod}} other_J$
$X_N \xleftarrow{\text{obj}} V \xrightarrow{\text{obj}} Y_N \xleftarrow{\text{mod}} other_J$
$X_N \xleftarrow{\text{obj}} handle_V \xrightarrow{\text{obj}} Y_N$
$X_N \xleftarrow{\text{obj}} V \xrightarrow{\text{obj}} Y_N$

Table 1: Examples of dependency patterns generated from the dependency path $screwdriver \xrightarrow{\text{conj}} hammer_N \xleftarrow{\text{obj}} handle_V \xrightarrow{\text{obj}} tool_N \xleftarrow{\text{mod}} other_J$

terns and patterns involving dependency information, are lexicalized patterns of co-occurrence. Due to data sparsity both approaches are generally characterized by a low recall. In this work we test strategies for improving the recall of systems leveraging patterns of co-occurrence created with dependency infomation, by using additional information extracted from corpora data. Our proposal consists in using all the dependency paths between two words up to three edges, instead of using only the shortest paths. We also introduce a generalization technique of the lexicalized patterns of co-occurrence based on part-of-speech information. We test these two strategies for classifying instances of five different lexical-semantic relations.

3. Problem description

In the present work we address the classification of unlabeled word pairs as instances of lexical-semantic relations. This means that given a set of target semantic relations $R = \{r_1, \ldots, r_n\}$, and a set of word pairs $W = \{(x, y)_1, \ldots, (x, y)_n\}$, the classification of relation instances consists in labeling each word pair $(x, y)_i$ with the relation $r_j \in R$ that holds between its members. Considering this, the output of the classification procedure is a set of tuples $((x, y)_i, r_j)$, where $(x, y)_i \in W$ and $r_j \in R$.

We approach the classification of relation instances as a supervised learner: given a set of target semantic relations $R = \{r_1, \ldots, r_n\}$, and a set of tuples $E = ((x, y)_i, r_i)$ of relation instances for each relation, our system learns features that are associated with each $r_i \in R$ and outputs a classifier. Then, given a new unlabeled pair $(x, y)_u$, the classifier decides if any of the relations $r_i \in R$ holds between x and y.

4. Pattern-based Classification ModEl

As aforementioned, in our work, we aim to analyze the impact of combining dependency information with a pattern generalization method based on part-of-speech information for the classification of instances of lexical-semantic relations.

Relying on the latent relation hypothesis (Turney, 2005), according to which pairs of words that co-occur in similar contexts tend to be instances of the same lexical-semantic relation, in our approach, word pairs are classified as instances of a semantic relation based on the context where its members co-occur. To do this, a set of features representing patterns of co-occurrence contexts is automatically created and used to create a vectorial representation for each word pair. A supervised system is trained to learn how to classify these word pairs based on the information in the feature vectors which represent their distributional behavior.

Further we refer to this system **Pa**ttern-based **C**lassification **E**ModEl (PaCE).

Feature Selection In our model, patterns of co-occurrrence are defined based on dependency relations between words as collected from corpus data. To extract these dependency relations, the input corpus was initially parsed using the Stanford Parser in the "collapsed dependency" format. Besides the dependency relations provided by the parser, we add a dependency relation vb_V connecting the subject with the object of a verb *V* in a given sentence.

We assume that all the contexts in which the members of any word pair (x, y) co-occur are likely to provide information regarding the relation that holds between them. Therefore, for each pair of words (x, y) that appears in the initial set of examples E, all sentences containing x and y are extracted. These sentences are then individually used to collect patterns of co-occurrence potentially indicating a semantic relation as follows: all the dependency paths between x and y up to three edges and containing only nouns, adjectives, verbs and adverbs are harvested from the dependency graph of each sentence; neighboring "window" nodes that are not included in the path but are connected to one of the members of the word pair are added to the path; each dependency path is transformed into a pattern of co-occurrence by unlexicalizing x and y, i.e. x and y are replaced with a slot which can be filled in by any word with the same part-of-speech.

For finding similarities between patterns of co-occurrence slightly differently lexicalized, each pattern is generalized using part-of-speech information: $2^{(n-2)}$ patterns are generated, n being the number of words on the initial dependency path, by iteratively replacing each word in the pattern with its part-of-speech. Window nodes are not submitted to this generalization procedure.

Finally, we filter out all the dependency patterns that do not co-occur with at least k unique example word pairs in our corpus. For the present experiments we set $k = 5$, following Snow et al. (2004).

For instance, from the dependency graph of the sentence "The students learned how to handle screwdrivers, hammers and other tools" shown in Figure 1, the path between *tool* and *hammer* generates the patterns shown in Table 1. The dependency relations inserted between the subject and the object of a verb in a sentence are represented using dashed red lines.

Word Pair Representation For classifying a word pair (x, y) as an instance of a lexical-semantic relation, the word pair is first represented as a vector of features. Each pattern of co-occurrence previously acquired is represented by a feature and the total set of patterns creates a feature space. The distributional behavior of each word pair is represented by a feature vector combining the features extracted from the various sentences in which the word pair co-occurs into one vector encoding in each position i the logarithm of the frequency of co-occurrence of x and y in the i^{th} pattern of

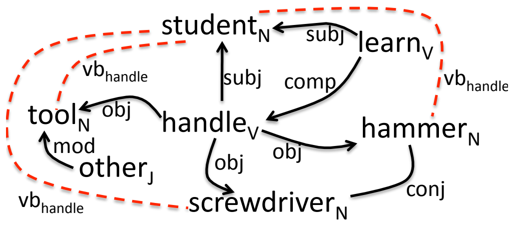


Figure 1: Dependency graph of the sentence "The students learned how to handle screwdrivers, hammers and other tools.". As mentioned, only words from the main part-of-speech are considered.

co-occurrence considered by the system[1].

Classification In our experiments word pairs are classified using an n-way multi-class SVM classifier with a radial basis function kernel (Platt, 1999). The system is trained by being provided with the feature vectors of word pairs from an initial small set of labeled relation instances. The trained SVM classifier is then used to determine whether an unlabeled word pair $(x, y)_u$ is an instance of a semantic relation $r_i \in R$.

As mentioned earlier, our approach relies on the assumption that any sentence in which a pair of words co-occurs is likely to provide information regarding the lexical semantic relation holding between them. Naturally, this does not necessarily hold for all the patterns of context acquired, as some of the features may not express any semantic relation of interest. For instance, the sentence *I can feel my fingers and close my hand.* contains the relation instance $((hand, finger), meronymy)$ but the context does not provide any information regarding the relation holding between *hand* and *finger*. We assume that the machine learning algorithm is able to discover which features are noisy, i.e. not informative for the task at hand, and to associate corresponding weights for minimizing errors in classification results.

5. Experiments

Our work addresses the classification of word pairs as instances of lexical-semantic relations. Systems developed to perform this task relying on the context of co-occurrence of words generally achieve good precision of classification but they typically score poorly on recall, as candidate word pairs must co-occur in the same type of context a sufficient number of times to provide enough distributional information to the system. The experiment presented in this paper has been designed to shed light on the impact of two strategies for potentially improving the recall of systems based on patterns of co-occurrence: the use of all the dependency paths between two words up to three edges instead of only the shortest ones; and the incorporation of a generalization strategy of the features used in classification based on part-of-speech information. These strategies are combined in the PaCE classification system (see Section 4.). For evaluating their impact in the classification of relation instances, we compare PaCE with three other systems.

PAIRCLASS The PairClass algorithm (Turney, 2008b) provides a state-of-the-art pattern-based approach for classifying the relationship between word pairs, which has performed well for many relation types. Using a set of seed pairs (x, y) for each relation, PairClass acquires a set of lexical patterns using the template

[0 to 1 words] x [0 to 3 words] y [0 to 1 words]

From the initial set of lexical patterns extracted from a corpus, additional patterns are generated by optionally generalizing each word to its part of speech. For N seed pairs, the most frequent kN patterns are retained. We follow Turney (2008b) and set $k = 20$. The patterns retained are then used as features to train an SVM classifier over the set of possible relation types. Finally, we underline that in the original experiments, PairClass was trained using a corpus of 5×10^{10} words, which is three orders of magnitude larger than the BNC corpus used in our experiments.

BASELINE As a baseline system for our experiments, we use a similar system to the one developed by Snow et al. (2004). This approach uses the shortest paths connecting two target words in the dependency graph of a sentence as features for the classification of word pairs. Besides, for each target word in the path, a neighboring window node (see Section 2) can be added to the dependency path. For a correct comparison, the baseline system uses the same classifier as the PaCE system.

BASELINE^{GEN} For highlighting the importance of the POS-generalization strategy incorporated in PaCE, we created a variant of the Baseline system that uses the same generalization strategy, but relies on the same features as the Baseline system.

PaCE^{LEX} For evaluating the specific contribution of considering all the dependency paths between two words instead of using only the shortest ones, and the importance of the POS-generalization strategy used in PaCE, we provide results obtained with PaCE^{LEX}, a variant of the PaCE system that uses as features only lexicalized dependency paths up to three edges.

PATTERNS For completeness, we also constructed a composite system using Hearst patterns (Hearst, 1992) for detecting hypernyms and co-hyponyms, and Bearland patterns (Berland and Charniak, 1999) for detecting meronyms. Classification is performed by measuring the frequency in corpus data of each relation pattern and then selecting the relation whose patterns occur more frequently[2].

[1] Preliminary experiments were run for finding the best performing way of representing the information in the vectors. We compared vectors in which occurrence information was represented with binary information, mutual information and the logarithm of the relative frequency. The best results were obtained using the logarithm of the frequency used in the experiments presented in this paper.

[2] This system classifies only hypernyms, co-hyponyms and

	Patterns			PairClass			Baseline			Baseline^{GEN}			PaCE^{LEX}			PaCE			UL
	P.	R.	F.	P.	R.	F.	P.	R.	F.	P.	R.	F.	P.	R.	F.	P.	R.	F.	R.
Attr				80.8	36.8	50.6	83.9	47.5	60.6	82.1	50.3	62.4	84.3	46.9	60.3	79.2	55.5	65.3	74
Coord	98.9	31.8	48.1	76.3	29.8	42.9	78	45.5	57.4	78.4	47.5	59.2	80.5	46.9	59.3	82	47.4	60.1	64.7
Act				73.5	46.4	56.9	78.7	64.7	71	79.1	65	71.3	78.3	64.6	70.8	78.2	65.3	71.2	83
Hypo	58.1	10.8	18.2	25.4	21.9	23.5	88	13.2	22.9	85.9	13.7	23.6	88.4	14.2	24.5	73	17	27.5	60
Mero	70.2	18.9	29.7	67.4	33.5	44.7	75	43.7	55.3	73	46.2	56.6	74.8	44	55.4	69.3	48.3	56.9	73.9
ALL	83.8	23.3	36.5	66.8	35.6	46.4	78.9	47.6	59.4	78.4	49.3	60.5	79.4	48	59.8	77.1	51.1	61.4	72.8

Table 2: Precision (P), Recall (R) and F-measure (F) obtained by the 5 systems and 2 variations considered, across all relation types. In the last column we present the upper-limit for recall (UL) (see Section 7).

Dataset To evaluate our system for the classification of lexical-semantic relations we used the BLESS dataset (Baroni and Lenci, 2011). This dataset provides relation instances for 200 concepts, covering 17 topical domains, e.g., tools or fruit. The relations considered are the taxonomic relations holding between nouns - hypernymy (*hammer, tool*), co-hyponymy, i.e., coordinates (*hammer, drill*), and meronymy (*hammer, handle*) -, but also attributes of nouns (*hammer, heavy*), and actions done by or to a noun (*hammer, beat*).

Corpora All systems were compared using distributional information collected in the British National Corpus (BNC), a balanced 100-million word corpus, parsed with the Stanford dependency parser.

Evaluation We compare system performance based on the scores obtained for precision (P), recall (R) and F1-measure (F). Precision is defined as the percentage of correct relation classifications of those made by a system; recall is defined as the percentage of relation instances in the dataset correctly classified by a system. The F1 measure is the harmonic mean of precision and recall. All the systems were tested using stratified 10-fold cross validation per domain. Results across all the domains are reported in Table 2.

6. Results

We compared the results obtained for each semantic relation and the results obtained across all the relations included in the BLESS dataset to assess the overall performance of each system. Overall, PaCE achieves statistically significant[3] better results across all the measures considered. It overcomes the PairClass system by 6.3 points in precision, 13 points in recall and 11.8 points in F-measure[4]. When compared with the Baseline, it scores 3.4 points higher in recall, losing 1.8 points in precision. The comparison between the Baseline and Baseline^{GEN} shows that by using a generalization strategy based on part-of-speech the recall increases by 1.7 points whereas no statistically significant difference is observed in precision scores. Comparing the PaCE system with PaCE^{LEX}, 3.5 points are gained in recall but the precision decreases by 2.3 points. However, according to the F-measure scores, which measures the balance between these two indicators of performance, PaCE is clearly the best performing system. The results for the Patterns system, calculated only for three relations (hypernymy, co-hyponymy and meronymy), show a high precision but a very low recall.

7. Discussion

All the systems relying on dependency information achieve better results than the PairClass system, which does not use this type of information. This comparison shows the impact this type of information has both on precision and recall across all the relations tested. The 8.1-point increase in precision achieved by the Baseline mirrors the greater reliability of patterns based on dependency information when compared against surface patterns. Additionally, the Baseline scores 12 points higher in recall, a difference that we attribute to the fact that these patterns go beyond the three word window used by Turney (2008b) and therefore provide a larger amount of the data available in corpus to the classification system.

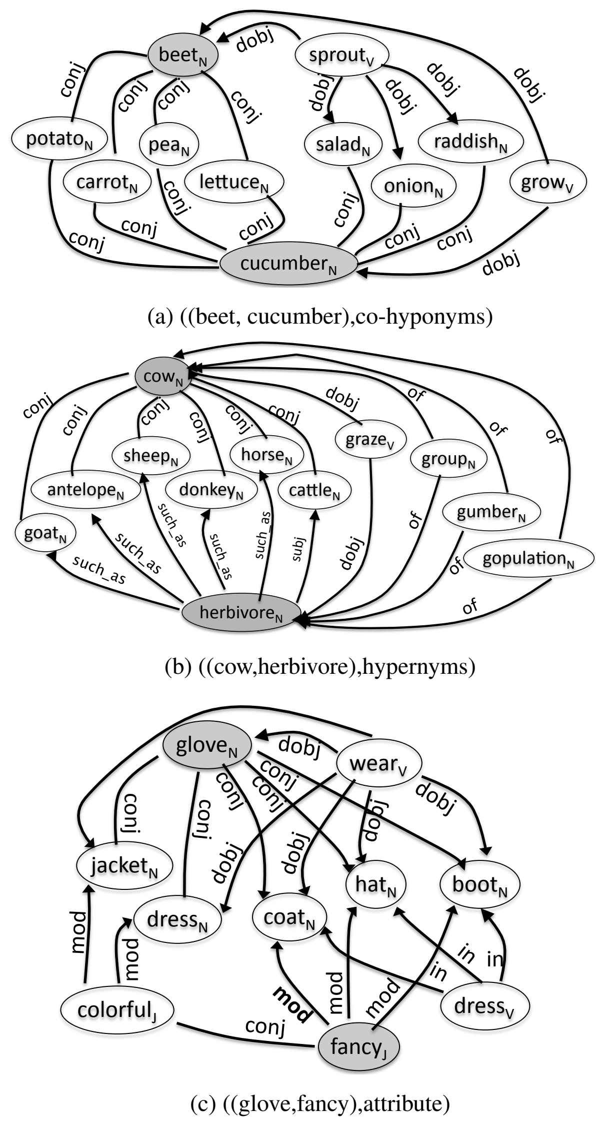
Comparing the PaCE^{LEX} system that uses all the dependency paths up to three edges with the Baseline system that uses only the shortest paths no statistically significant differences are observed in the results. Therefore, we can conclude that even when only the shortest dependency paths are considered, the most important patterns of co-occurrence are already gathered.

The positive impact of combining information on all dependency paths up to three edges with the POS-generalization strategy in the recall scores, as made apparent by the scores of the PaCE system, has nonetheless to be underlined. In fact, comparing Baseline^{GEN} with the Baseline system, the importance of this generalization strategy is highlighted. As expected, generalizing the patterns of co-occurrence with part-of-speech information improves recall by 1.7 points. However, it is when all the dependency paths up to three edges are used in combination with the POS-generalization strategy that recall further improves by 3.5 points, although the features used are apparently less reliable as the precision drops 2.3 points.

meronyms because no standard manually-designed patterns exist in the literature for the two remaining lexical-semantic relations considered in our experiments.

[3] Statistical significance was calculated using Student's t-test with a 95-percent confidence interval.

[4] Furthermore, it should be noted that the approach used in the PairClass system, in contrast with all the other systems considered, does not seem to work consistently for all types of semantic relations (see the poor performance of the system for the hypernymy relation in Table 2). Although further research would have to be conducted to support any analysis of these contrasts, the results obtained seem to suggest that the more asymmetric a semantic relation is, the more poorly this system performs.



(a) ((beet, cucumber),co-hyponyms)

(b) ((cow,herbivore),hypernyms)

(c) ((glove,fancy),attribute)

Figure 2: Local networks of two elements unrelated in corpus

However, a major point of discussion raised by the experiments conducted in the context of the work presented in this paper is the main limitation of pattern-based systems for the classification of semantic relation instances imposed by the fact that word pairs are classified based on evidence regarding their co-occurrence in the same sentence. This way, the number of word pairs occurring in the same sentence in the input corpus constitutes the real upper-limit (UL) of these approaches (see the last column in Table 2) in terms of recall. In our experiments, only 72.8% of the relation instances included in BLESS co-occur at least once in the same sentence in the BNC corpus. Being so, all the remaining candidate pairs cannot be classified due to a total lack of information regarding their joint distributional behavior. Moreover, this is not a corpus-specific limitation, since due to the zipfian distribution of words, in any corpus of any size there will always be word pairs that will not co-occur in corpus to provide enough information to classifiers or to any automatic system based on distributional information. For instance, out of the word pairs that are misclassified by PaCE, 60% have a low-frequency of co-occurrence in the same sentence.

This problem has been addressed in the literature by combining approaches based on patterns extracted from co-occurrence contexts with semantic similarity between words (Turney, 2006b; Herdadelen and Baroni, 2009). The results obtained in solving analogies from the SAT test[5] show that the best performances were achieved by systems highly dependent on a very large corpus of ∼50Gb (Turney, 2006b; Turney, 2013; Turney, 2006a; Turney, 2008b). In order to scale down the dimensions of the input corpus, new techniques are necessary to go beyond the aforementioned upper-limit imposed by the amount of co-occurrences in the same sentence observed in a given corpus and acquire more information to identify the relation holding between pairs of words that co-occur very infrequently or not at all in the same sentence in a given corpus, although they are related.

In order to get some insight on possible strategies for tackling this limitation of pattern-based classification systems for relation instances, we ran an empirical error analysis of PaCE results, in which we observed that 13% of the dataset does not co-occur in the same sentence. Yet, the members of these word pairs share a significant number of co-occurring words in the same type of dependency relation. In Figure 2 we present instances of co-hyponymy, hypernymy and attribute relations from our dataset that do not co-occur in the corpus, although they are densely linked when "bridging words" are considered.

$Beet_N$ and $cucumber_N$ are two co-hyponyms, both hyponyms of vegetable, which do not co-occur in the same sentence in our corpus. However, both occur in a relation of coordination with other hyponyms of vegetable such as $potato_N$, $carrot_N$, $lettuce_N$ and pea_N. Additionally, both target nouns occur as the direct object of the verb $grow_V$, while only $beet_N$ occurs as direct object of the verb $sprout_V$. However, $salad_N$, $onion_N$ and $raddish_N$ do occur as a direct object of the verb $sprout_V$ and they occur in coordination with $cucumber_N$.

Regarding the relation of hypernymy, cow_N and $herbivore_N$ are an example of a word pair from the domain of *ground mammals* which instantiate this relation but does not co-occur in the same sentence in BNC, although both nouns occur in a conjunction with other herbivores such as $antelope_N$, $goat_N$, $sheep_N$, $donkey_N$, $horse_N$ and $cattle_N$. This set of words co-occurs with $herbivore_N$ in the pattern *herbivore such as X* or *herbivore is a X*. Also, both *cow* and *herbivore* occur with phrases like *group of X*, *number of X*, *population of X* and *herd of X*, as well as as objects of the verb $graze_V$.

Finally, in our dataset, $fancy_J$ is an attribute of $glove_N$ but these words do not co-occur in the same sentence in BNC. In our corpus, however, $fancy_J$ is the modifier of $boot_N$, $coat_N$, $dress_N$, hat_N, $jacket_N$ and tie_N, among many others, which are words that co-occur in conjunction with the word $glove_N$. Additionally, all of these words are objects of the verb *to* $wear_V$ and they occur in the phrase *dressed in X*, just like $glove_N$ does. Although $fancy_J$ and $glove_N$ do not co-occur, $fancy_J$ co-occurs

[5] http://aclweb.org/aclwiki/index.php?title=SAT_Analogy_Questions_(State_of_the_art)

with $colorful_J$, which is an adjectival modifier of $jacket_N$ and $dress_N$, two words that occur in a coordination relation with $glove_N$.

Therefore, for detecting new relation instances, valuable information can be extracted from parallel sentences separately containing occurrences of each individual member of a candidate pair and where the members of the pair share other co-occurring words. The vectorial representation of each individual target word captures this type of information but it is unclear how to use this information to assess whether a given lexical-semantic relation holds or not between a candidate word pair. And yet, as made apparent by the examples above, dependency relations between individual target words and shared third party co-occurring words may be a valuable indication of the relation holding between a word pair.

In future work we plan to investigate possible strategies to combine this type of shared dependency information between the individual members of a candidate word pair.

8. Conclusions

The present work addresses the impact of using all the dependency paths up to three edges, instead of only the shortest one, in combination with a generalization strategy using part-of-speech information in the classification of instances of lexical-semantic relations. The isolated use of all the dependency paths up to three edges has not resulted in any statistically significant improvements when compared with using only the shortest paths. POS-generalized patterns of co-occurrence, however, are able to correctly classify more candidate word pairs. But it is the combination of these two strategies that yield the highest increase in recall, 3.5 points, although this apparently results in a loss of precision of the patterns of co-occurrence, as made apparent by the drop in precision by 2.3 points.

Finally, running an error analysis of the results obtained made apparent the need for new techniques to go beyond sentence boundaries when harvesting information in pattern-based models for the classification of relation instances.

9. Acknowledgements

The authors gratefully acknowledge the support of the CLARA project (EU-7FP-ITN-238405), of the SKATER project (Ministerio de Economa y Competitividad, TIN2012-38584-C06-05) and of Fundação para a Ciência e a Tecnologia (FCT) post-doctoral fellowship SFRH/BPD/79900/2011.

10. References

Baroni, M. and Lenci, A. (2011). How we blessed distributional semantic evaluation. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of ACL*, pages 57–64. Association for Computational Linguistics.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250. ACM.

Davidov, D. and Rappoport, A. (2006). Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. In *Proceedings of COLING-ACL*, pages 297–304.

Girju, R., Badulescu, A., and Moldovan, D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Girju, R., Badulescu, A., and Moldovan, D. I. (2006). Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135.

Girju, R. (2003). Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering - Volume 12*, MultiSumQA '03, pages 76–83, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING*, pages 539–545.

Herdadelen, A. and Baroni, M. (2009). Bagpack: A general framework to represent semantic relations. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 33–40. Association for Computational Linguistics.

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., and Weld, D. S. (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 541–550. Association for Computational Linguistics.

Lin, D., Zhao, S., Qin, L., and Zhou, M. (2003). Identifying synonyms among distributionally similar words. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, pages 1492–1493, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of ACL-CONLL*, pages 1003–1011. Association for Computational Linguistics.

Pasca, M. and Harabagiu, S. M. (2001). The informative role of WordNet in open-domain question answering. In *Proceedings of NAACL-01 Workshop on WordNet and Other Lexical Resources*, pages 138–143.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In Schölkopf, B., Burges, C. J. C., and Smola, A. J., editors, *Advances in kernel methods*, pages 185–208. MIT Press, Cambridge, MA, USA.

Ravichandran, D. and Hovy, E. H. (2002). Learning surface text patterns for a question answering system. In *Proceedings of ACL*, pages 41–47.

Snow, R., Jurafsky, D., and Ng, A. Y. (2004). Learning syntactic patterns for automatic hypernym discovery. In *Proceedings of NIPS*.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA. ACM.

Surdeanu, M., Tibshirani, J., Nallapati, R., and Manning, C. D. (2012). Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465. Association for Computational Linguistics.

Turney, P. (2005). Measuring semantic similarity by latent relational analysis. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05)*.

Turney, P. D. (2006a). Expressing implicit semantic relations without supervision. In *Proceedings of COLING-ACL*, pages 313–320. Association for Computational Linguistics.

Turney, P. D. (2006b). Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Turney, P. D. (2008a). The latent relation mapping engine: Algorithm and experiments. *J. Artif. Intell. Res.(JAIR)*, 33:615–655.

Turney, P. D. (2008b). A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of COLING*, pages 905–912.

Turney, P. D. (2013). Distributional semantics beyond words: Supervised learning of analogy and paraphrase. *arXiv preprint arXiv:1310.5042*.

Wu, F. and Weld, D. S. (2010). Open information extraction using wikipedia. In *Proceedings of ACL*, pages 118–127.