Finding Word Substitutions Using a Distributional Similarity Baseline and Immediate Context Overlap

Aurelie Herbelot
University of Cambridge
Computer Laboratory
J.J. Thompson Avenue
Cambridge
ah433@cam.ac.uk

Abstract

This paper deals with the task of finding generally applicable substitutions for a given input term. We show that the output of a distributional similarity system baseline can be filtered to obtain terms that are not simply similar but frequently substitutable. Our filter relies on the fact that when two terms are in a common entailment relation, it should be possible to substitute one for the other in their most frequent surface contexts. Using the Google 5-gram corpus to find such characteristic contexts, we show that for the given task, our filter improves the precision of a distributional similarity system from 41% to 56% on a test set comprising common transitive verbs.

1 Introduction

This paper looks at the task of finding word substitutions for simple statements in the context of KB querying. Let us assume that we have a knowledge base made of statements of the type 'subject – verb – object':

1. Bank of America – acquire – Merrill Lynch
2. Lloyd's – buy – HBOS
3. Iceland – nationalise – Kaupthing

Let us also assume a simple querying facility, where the user can enter a word and be presented with all statements containing that word, in a typical search engine fashion. If we want to return all acquisition events present in the knowledge base above (as opposed to nationalisation events), we might search for 'acquire'. This will return the first statement (about the acquisition of Merrill Lynch) but not the second statement about HBOS. Ideally, we would like a system able to generate words similar to our query, so that a statement containing the verb 'buy' gets returned when we search for 'acquire'.

This problem is closely related to the clustering of semantically similar terms, which has received much attention in the literature. Systems that perform such clustering usually do so under the assumption of *distributional similarity* (Harris, 1954) which state that two words appearing in similar contexts will be close in meaning. This observation is statistically useful and has contributed to successful systems within two approaches: the pattern-based approach and the feature vector approach (we describe those two approaches in the next section). The definition of similarity used by those systems is fairly wide, however. Typically, a query on the verb 'produce' will return verbs such as 'export', 'import' or 'sell', for instance (see DIRT demo from `http://demo.patrickpantel.com/Content/LexSem/paraphrase.htm`, Lin and Pantel, 2001.)

This fairly wide notion of similarity is not fully appropriate for our word substitutions task: although cats and dogs are similar types of entities, querying a knowledge base for 'cat' shouldn't return statements about dogs; statements about Siamese, however, should be acceptable. So, following Dagan and Glickman (2004), we refine our concept of similarity as that of *entailment*, defined here as the relation whereby the meaning of a word w_1 is 'included' in the meaning of word w_2 (practically speaking, we assume that the 'meaning' of a word is represented by the contexts in which it appears and require that if w_1 entails w_2, the contexts of w_2 should be a subset of the contexts of w_1). Given an input term w, we therefore attempt to extract words which either entail or are entailed by w. (We do not extract directionality at this stage.)

The definition of entailment usually implies that an entailing word must be *substitutable* for the entailed one, in *some* contexts at least. Here, we consider word substitution queries in cases where no additional contextual information is given, so we cannot assume that possible, but rare, substitutions will fit the query intended by the user ('believe' correctly entails 'buy' in some cases but we can be reasonably sure that the query 'buy' is meant in the 'purchase' sense.) We thus require that our output will fit the *most* common contexts. For instance, given the query 'kill', we want to return 'murder' but not 'stop'. Given 'produce', we want to return both 'release' and 'generate' but not 'fabricate' or 'hatch'.[1] Taking this into account, we generally define *substitutability* as the ability of a word to replace another one in a given sentence without changing the meaning or acceptability of the sentence, and this *in the most frequent cases*. (By *acceptability*, we mean whether the sentence is likely to be uttered by a native speaker of the language under consideration.)

In order to achieve both entailment and general substitutability, we propose to filter the output of a conventional distributional similarity system using a check for lexical substitutability in frequent contexts. The idea of the filter relies on the observation that entailing words tend to share more frequent immediate contexts than just related ones. For instance, when looking at the top 200 most frequent Google 3-gram contexts (Brants and Franz, 2006) appearing after the terms 'kill', 'murder' and 'abduct', we find that 'kill' and 'murder' share 54 while 'kill' and 'abduct' only share 2, giving us the indication that as far as usage is concerned, 'murder' is closer to 'kill' than 'abduct'. Additionally, context frequency provides a way to identify substitutability for the most common uses of the word, as required.

In what follows, we briefly present related work, and introduce our corpus and algorithm, including a discussion of our 'immediate context overlap' filter. We then review the results of an experiment on the extraction of entailment pairs for 30 input verbs.

[1] In fact, we argue that even in systems where context is available, searching for all entailing words is not necessary an advantage: consider the query 'What does Dole produce?' to a search engine. The verb 'fabricate' entails 'produce' in the correct sense of the word, but because of its own polysemy, and unless an expensive layer of WSD is added to the system, it will return sentences such as 'Dole fabricated stories about her opponent', which is clearly not the information that the user was looking for.

2 Previous Work

2.1 Distributional Similarity

2.1.1 Principles

Systems using distributional similarity usually fall under two approaches:

1. The pattern-based approach (e.g. Ravichadran and Hovy, 2002). The most significant contexts for an input seed are extracted as features and those features used to discover words related to the input (under the assumption that words appearing in *at least one* significant context are similar to the seed word). There is also a non-distributional strand of this approach: it uses Hearst-like patterns (Hearst, 1992) which are supposed to indicate the presence of two terms in a certain relation - most often hyponymy or meronymy (see Chklovski and Pantel, 2004).

2. The feature vector approach (e.g. Lin and Pantel, 2001). This method fully embraces the definition of distributional similarity by making the assumption that two words appearing in similar *sets* of features must be related.

2.1.2 Limitations

The problems of the distributional similarity assumption are well-known: the facts that 'a bank lends money' and 'Smith's brother lent him money' do not imply that banks and brothers are similar entities. This effect becomes particularly evident in cases where antonyms are returned by the system; in those cases, a very high distributional similarity actually corresponds to opposite meanings. Producing an output ranked according to distributional similarity scores (weeding out anything under a certain threshold) is therefore not sufficient to retain good precisions for many tasks. Some work has thus focused on a re-ranking strategies (see Geffet and Dagan, 2004 and Geffet and Dagan, 2005, who improve the output of a distributional similarity system for an entailment task using a web-based feature inclusion check, and comment that their filtering produces better outputs than cutting off the similarity pairs with the lowest ranking.)

2.2 Extraction Systems

Prominent entailment rule acquisition systems include DIRT (Lin and Pantel, 2001), which uses distributional similarity on a 1 GB corpus to identify semantically similar words and expressions, and TEASE (Szpektor *et al.*, 2004), which extracts entailment relations from the web for a given word by computing characteristic contexts for that word.

Recently, systems that combine both pattern-based and feature vector approaches have also been presented. Lin *et al.* (2003) and Pantel and Ravichandran (2004) have proposed to classify the output of systems based on feature vectors using lexico-syntactic patterns, respectively in order to remove antonyms from a related words list and to name clusters of related terms.

Even more related to our work, Mirkin *et al.* (2006) integrate both approaches by constructing features for the output of both a pattern-based and a vector-based systems, and by filtering incorrect entries with a supervised SVM classifier. (The pattern-based approach uses a set of manually-constructed patterns applied to a web search.)

In the same vein, Geffet and Dagan (2005) filter the result of a pattern-based system using feature vectors. They get their features out of an 18 million word corpus augmented by a web search. Their idea is that for any pair of potentially similar words, the features of the entailed one should comprise all the features of the entailing one.

The main difference between our work and the last two quoted papers is that we add a new layer of verification: we extract pairs of verbs using automatically derived semantic patterns, perform a first stage of filtering using the semantic signatures of each word and apply a final stage of filtering relying on surface substitutability, which we name 'immediate context overlap' method. We also experiment with a smaller size corpus to produce our distributional similarity baseline (a subset of Wikipedia) in an attempt to show that a good semantic parse and adequate filtering can provide reasonable performance even on domains where data is sparse. Our method does not need manually constructed patterns or supervised classifier training.

2.3 Evaluation

The evaluation of KB or ontology extraction systems is typically done by presenting human judges with a subset of extracted data and asking them to annotate it according to certain correctness criteria. For entailment systems, the annotation usually relies on two tests: whether the meaning of one word entails the other one in some senses of those words, and whether the judges can come up with contexts in which the words are directly substitutable. Szpektor *et al.* (2007) point out the difficulties in applying those criteria. They note the low inter-annotator agreements obtained in previous studies and propose a new evaluation method based on precise judgement questions applied to a set of relevant contexts. Using their methods, they evaluate the DIRT (Lin and Pantel, 2001) and TEASE (Szpektor *et al.*, 2004) algorithms and obtain upper bound precisions of 44% and 38% respectively on 646 entailment rules for 30 transitive verbs. We follow here their methodology to check the results obtained via the traditional annotation.

3 The Data

The corpus used for our distributional similarity baseline consists of a subset of Wikipedia totalling 500 MB in size, parsed first with RASP2 (Briscoe *et al.*, 2006) and then into a Robust Minimal Recursion Semantics form (RMRS, Copestake, 2004) using a RASP-to-RMRS converter. The RMRS representation consists of trees (or tree fragments when a complete parse is not possible) which comprise, for each phrase in the sentence, a semantic head and its arguments. For instance, in the sentence 'Lloyd's rescues failing bank', three subtrees can be extracted:

`lemma:rescue arg:ARG1 var:Lloyd's`

which indicates that 'Lloyd's' is subject of the head 'rescue',

`lemma:rescue arg:ARG2 var:bank`

which indicates that 'bank' is object of the head 'rescue', and

`lemma:failing arg:ARG1 var:bank`

which indicates that the argument of 'failing' is 'bank'.

Note that any tree can be transformed into a feature for a particular lexical item by replacing the slot containing the word with a hole: `lemma:rescue arg:ARG2 var:bank` becomes `lemma:hole_ arg:ARG2 var:bank`, a potentially characteristic context for 'rescue'.

All the experiments reported in this paper concern transitive verbs. In order to speed up processing, we reduced the RMRS corpus to a

list of relations with a verbal head and at least two arguments: `lemma:verb-query arg:ARG1 var:subject arg:ARG2 var:object`. Note that we did not force noun phrases in the second argument of the relations and for instance, the verb 'say' was both considered as taking a noun or a clause as second argument ('to say a word', 'to say that the word is...').

4 A Baseline

We describe here our baseline, a system based on distributional similarity.

4.1 Step 1 - Pattern-Based Pair Extraction

The first step of our algorithm uses a pattern-based approach to get a list of potential entailing pairs. For each word w presented to the system, we extract all semantic patterns containing w. Those semantic patterns are RMRS subtrees consisting of a semantic head and its children (see Section 3). We then calculate the Pointwise Mutual Information between each pattern p and w:

$$pmi(p,w) = \log\left(\frac{P\left(p,w\right)}{P\left(p\right)P\left(w\right)}\right) \qquad (1)$$

where $P(p)$ and $P(w)$ are the probabilities of occurrence of the pattern and the instance respectively and $P(p,w)$ is the probability that they appear together.

PMI is known to have a bias towards less frequent events. In order to counterbalance that bias, we apply a simple logarithm function to the results as a discount:

$$d = \log\left(c_{wp}+1\right) \qquad (2)$$

where c_{wp} is the cooccurrence count of an instance and a pattern.

We multiply the original PMI value by this discount to find the final PMI. We then select the n patterns with highest PMIs and use them as relevant semantic contexts to find all terms t that also appear in those contexts. The result of this step is a list of potential entailment relations, $w - t_1$ ... $w - t_x$ (we do not know the direction of the entailment).

4.2 Step 2 - Feature vector Comparison

This step takes the output of the pattern-based extraction and applies a first filter to the potential entailment pairs. The filter relies on the idea that two words that are similar will have similar feature vectors (see Geffet and Dagan, 2005). We define here the feature vector of word w as the list of semantic features containing w, together with the PMI of each feature in relation to w as a weight. For each pair of words $(w1, w2)$ we extract the feature vectors of both $w1$ and $w2$ and calculate their similarity using the measure of Lin (1998). Pairs with a similarity under a certain threshold are weeded out. (We use 0.007 in our experiments – the value was found by comparing precisions for various thresholds in a set of initial experiments.)

As a check of how the Lin measure performed on our Wikipedia subset using RMRS features, we reproduced the Miller and Charles experiment (1991) which consists in asking humans to rate the similarity of 30 noun pairs. The experiment is a standard test for semantic similarity systems (see Jarmasz and Szpakowicz, 2003; Lin, 1998; Resnik, 1995 and Hirst and St Onge, 1998 amongst others). The correlations obtained by previous systems range between the high 0.6 and the high 0.8. Those systems rely on edge counting using manually-created resources such as WordNet and the Roget's Thesaurus. We are not actually aware of results obtained on totally automated systems (apart from a baseline computed by Strube and Ponzetto, 2006, using Google hits, which return a correlation of 0.26.)

Applying our feature vector step to the Miller and Charles pairs, we get a correlation of 0.38, way below the edge-counting systems. It turns out, however, that this low result is at least partially due to data sparsity: when ignoring the pairs containing at least one word with frequency under 200 (8 of them, which means ending up with 22 pairs left out of the initial 30), the correlation goes up to 0.69. This is in line with the edge-counting systems and shows that our baseline system produces a decent approximation of human performance, as long as enough data is supplied. [2]

Two issues remain, though. First, fine-grained results cannot be obtained over a general corpus: we note that the pairs 'coast-forest' and 'coast-hill' get very similar scores using distributional similarity while the latter is ranked twice as high as the former by humans. Secondly, distribu-

[2] It seems then that in order to maintain precision to a higher level on our corpus, we could simply disregard pairs with low-frequency words. (We decided here, however, that this would be unacceptable from the point of view of recall and did not attempt to do so.)

tional methods promise to identify 'semantically similar' words, as do the Miller and Charles experiment and edge-counting systems. However, as pointed out in the introduction, there is still a gap between general similarity and entailment: 'coast' and 'hill' are indeed similar in some way but never substitutable. Our baseline is therefore constrained by a theoretical problem that further modules must solve.

5 Immediate Context Overlap

Our immediate context overlap module acts as a filter for the system described as our baseline. The idea is that, out of all pairs of 'similar' words, we want to find those that express entailment in at least one direction. So for instance, given the pairs 'kill – murder' and 'kill – abduct', we would like to keep the former and filter the latter out. We can roughly explain why the second pair is not acceptable by saying that, although the semantics of the two words are close (they are both about an act of violence conducted against somebody), they are not substitutable in a given sentence.

To satisfy substitutability, we generally specify that if $w1$ entails $w2$, then there should be surface contexts where $w2$ can replace $w1$, with the substitution still producing an acceptable utterance (see our definition of *acceptability* in the introduction). We further suggest that if one word can substitute the other in frequent immediate contexts, we have the basis to believe that entailment is possible in at least one common sense of the words – while if substitution is impossible or rare, we can doubt the presence of an entailment relation, at least in common senses of the terms. This can be made clearer with an example. We show in Table 1 some of the most frequent trigrams to appear after the verbs 'to kill', 'to murder' and 'to abduct' (those trigrams were collected from the Google 5-gram corpus.) It is immediately noticeable that some contexts are not transferable from one term to the other: phrases such as 'to murder and forcibly recruit someone', or 'to abduct cancer cells' are impossible – or at least unconventional. We also show in italic some common immediate contexts between the three words. As pointed out in the introduction, when looking at the top 200 most frequent contexts for each term, we find that 'kill' and 'murder' share 54 while 'kill' and 'abduct' only share 2, giving us the indication that as far as usage is concerned, 'murder' is closer to 'kill' than 'abduct'. Furthermore, by looking at frequency of occurrence, we partly answer our need to find substitutions that work in very frequent sentences of the language.

The Google 5-gram corpus gives the frequency of each of its n-grams, allowing us to check substitutability on the 5-grams with highest occurrence counts for each potential entailment pair returned by our baseline. For each pair $(w1, w2)$ we select the m most frequent contexts for both $w1$ and $w2$ and simply count the overlap between both lists. If there is any overlap, we keep the pair; if the overlap is 0, we weed it out (the low threshold helps our recall to remain acceptable). We experiment with left and right contexts, i.e. with the query term at the beginning and the end of the n-gram, and with various combinations (see Section 6).

6 Results

The results in this section are produced by randomly selecting 30 transitive verbs out of the 500 most frequent in our Wikipedia corpus and using our system to extract non-directional entailment pairs for those verbs, following a similar experiment by Szpektor *et al.* (2007). We use a list of $n = 30$ features in Step 1 of the baseline. We evaluate the results by first annotating them according to a broad definition of entailment: if the annotator can think of any context where one word of the pair could replace the other, preserving surface form and semantics, then the two words are in an entailment relation. (Note again that we do not consider the directionality of entailment at this stage.) We then re-evaluate our best score using the Szpektor *et al.* method (2007), which we think is more suited for checking true substitutability. [3]

The baseline described in Section 4 produces 301 unique pairs, 124 of which we judge correct using our broad entailment definition, yielding a precision of 41%. The average number of relations extracted for each input term is thus 4.1.

Tables 2 and 3 show our results at the end of the immediate context overlap step. Table 2 report results using the $m = 50$ most frequent contexts for each word in the pair while Table 3 uses an expanded list of 200 contexts. Precision is the

[3] Although no direct comparison with the works of Szpektor *et al.* or Lin and Pantel is provided in this paper, we are in the process of evaluating our results against the TEASE output (available at `http://www.cs.biu.ac.il/~szpekti/TEASE_collection.zip`) through a web-based annotation task.

Table 1: Immediate Contexts for 'kill', 'murder' and 'abduct'

kill	murder	abduct
two birds with	babies that life	her and make
cancer cells and	*his wife and*	an innocent man
a mocking bird	thousands of innocent	unsuspecting people and
or die for	*women and children*	suspects in foreign
or be killed	her husband and	a young girl
another human being	*in the name*	and forcibly recruit
thousands of people	in connection with	a teenage girl
in the name	*another human being*	and kill her
his wife and	tens of thousands	a child from
members of the	the royal family	*women and children*

number of correct relations amongst all those returned. Recall is calculated with regard to the 124 pairs judged correct at the end of the previous step (i.e., this is not true recall but recall relative to the baseline results.)

We experimented with six different set-ups:

1- right context: the four words following the query term are used as context

2- left context: the four words preceding the query term are used as context

3- right and left contexts: the best contexts (those with highest frequencies) are selected out of the concatenation of both right and left context lists

4- concatenation: the concatenation of the results obtained from 1 and 2

5- inclusion: the inclusion set of the results from 1 and 2, that is, the pairs judged correct by *both* the right context and left context methods.

6- right context with 'to': identical to 1 but the 5-gram is required to start with 'to'. This ensures that only the verb form of the query term is considered but has the disadvantage of effectively transforming 5-grams into 4-grams.

Our best overall results comes from using 50 immediate contexts starting with 'to', right context only: we obtain 56% precision on a recall of 85% calculated on the results of the previous step.

Table 2: Results using 50 immediate contexts

Context Used	Precision	Recall	F	Returned	Correct
Left	48%	63%	54%	164	78
Right	62%	26%	36%	52	32
Left and Right	53%	52%	52%	122	65
Concatenation	48%	70%	57%	181	87
Inclusion	67%	19%	30%	36	24
Right + 'to'	56%	85%	68%	187	105

Table 3: Results using 200 immediate contexts

Context Used	Precision	Recall	F	Returned	Correct
Left	44%	86%	58%	244	107
Right	54%	60%	57%	137	74
Left and Right	46%	85%	60%	228	105
Concatenation	44%	92%	60%	260	114
Inclusion	55%	53%	54%	121	66
Right + 'to'	48%	97%	64%	248	120

6.1 Instance-Based Evaluation

We then recalculate our best precision following the method introduced in Szpektor *et al.* (2007). This approach consists in extracting, for each potential entailment relation X-$verb_1$-Y $\Rightarrow$ X-$verb_2$-Y, 15 sentences in which $verb1$ appears and ask annotators to provide answers to three questions:

1. Is the left-hand side of the relation entailed by the sentence? If so...

2. When replacing $verb_1$ with $verb_2$, is the sentence still likely in English? If so...

3. Does the sentence with $verb_1$ entail the sentence with $verb_2$?

We show in Table 4 some potential annotations at various stages of the process.

For each pair, Szpektor *et al.* then calculate a lower-bound precision as

$$P_{lb} = \frac{n_{Entailed}}{n_{LeftHandEntailed}} \quad (3)$$

where $n_{Entailed}$ is the number of entailed sentence pairs (the annotator has answered 'yes' to the third question) and $n_{LeftHandEntailed}$ is the number of sentences where the left-hand relation is entailed (the annotator has answered 'yes' to the first question). They also calculate an upper-bound precision as

$$P_{ub} = \frac{n_{Entailed}}{n_{Acceptable}} \quad (4)$$

where $n_{Acceptable}$ is the number of acceptable $verb_2$ sentences (the annotator has answered 'yes' to the second question). A pair is deemed to contain an entailment relation if the precision for that particular pair is over 80%.

The authors comment that a large proportion of extracted sentences lead to a 'left-hand side not entailed' answer. In order to counteract that effect, we only extract sentences without modals or negation from our Wikipedia corpus and consequently only require 10 sentences per relation (only 11% of our sentences have a 'non-entailed' left-hand side relation against 43% for Szpektor *et al.*).

We obtain an upper bound precision of 52%, which is slightly lower than the one initially calculated using our broad definition of entailment, showing that the more stringent evaluation is useful when checking for general substitutability in the returned pairs. When we calculate the lower bound precision, however, we obtain a low 10% precision due to the large number of sentences judged as 'unlikely English sentences' after substitution (they amount to 33% of all examples with a left-hand side judged 'entailed'). This result illustrates the need for a module able to check sentence acceptability when applying the system to true substitution tasks. Fortunately, as we explain in the next section, it also takes into account requirements that are only necessary for generation tasks, and are therefore irrelevant to our querying task.

7 Discussion

Our main result is that the immediate context overlap step dramatically increases our precision (from 41% to 56%), showing that a more stringent notion of similarity can be achieved when adequately filtering the output of a distributional similarity system. However, it also turns out that looking at the most frequent contexts of the word to substitute does not fully solve the issue of surface *acceptability* (leading to a high number of 'right-hand side not entailed' annotations). We argue, though, that the issue of producing an acceptable English sentence is a generation problem separate from the extraction task. Some systems, in fact, are dedicated to related problems, such as identifying whether the senses of two synonyms are the same in a particular lexical context (see Dagan *et al.*, 2006). As far as our needs are concerned in the task of KB querying, we only require accurate searching capabilities as opposed to generational capabilities: the expansion of search terms to include impossible strings is not a problem in terms of result.

Looking at the immediate context overlaps returned for each pair by the system, we find that the overlap (the similarity) can be situated at various linguistic layers:

- in the semantics of the verb's object: 'a new album' is something that one would frequently 'record' or 'release'. The phrase boosts the similarity score between 'record' and 'release' in their music sense.
- in the clausal information of the right context: a context starting with a clause introduced by 'that' is likely to be preceded by a verb expressing cognition or discourse. The tri-gram 'that there is' increases the similarity of pairs such as 'say - argue'.
- in the prepositional information of the right context: 'about' is the preposition of choice after cognition verbs such as 'think' or 'wonder'. The context 'about the future' helps the score of the pair 'think - speculate' in the cognitive sense (note that 'speculate' in a financial sense would take the preposition 'on'.)

Some examples of overlaps are shown in Table 5.

We also note that the system returns a fair proportion of vacuous contexts such as 'one of the' or

Table 4: Annotation Examples Following the Szpektor *et al.* Method

Word Pair	Sentence	Question 1	Question 2	Question 3
acquire – buy	Lloyds acquires HBOS	yes	yes (Lloyds buys HBOS)	yes
acquire – praise	Lloyds acquires HBOS	yes	yes (Lloyds praises HBOS)	no
acquire – spend	Lloyds acquires HBOS	yes	no (*Lloyds spends HBOS)	–
acquire – buy	Lloyds may acquire HBOS	no	–	–

Table 5: Sample of Immediate Context Overlaps

think – speculate	say – claim	describe – characterise
about the future	that it is	the nature of
about what the	that there is	the effects of
about how the	that it was	it as a
	that they were	the effect of
	that they have	the role of
	that it has	the quality of
		the impact of
		the dynamics of

Table 6: Sample of Extracted Pairs

bring – attract	make - earn
*call – form	*name - delegate
change – alter	offer - provide
create – generate	*perform - discharge
describe – characterise	produce – release
develop – generate	record – count
*do – behave	*release – announce
feature – boast	*remain – comprise
*find – indicate	require – demand
follow – adopt	say – claim
*grow – contract	tell – assure
*increase - decline	think – believe
leave - abandon	*use – abandon

'part of the' which contribute to the score of many pairs. Our precision would probably benefit from excluding such contexts.

We note that as expected, using a larger set of contexts leads to better recall and decreased precision. The best precision is obtained by returning the inclusion set of both left and right contexts results, but at a high cost in recall. Interestingly, we find that the right context of the verb is far more telling than the left one (potentially, objects are more important than subjects). This is in line with results reported by Alfonseca and Manandhar (2002).

Our best results yield an average of 3.4 relations for each input term. It is in the range reported by the authors of the TEASE system (Szpektor *et al.*, 2004) but well below the extrapolated figures of over 20 relations in Szpektor *et al.*, 2007. We point out, however, that we only search for single word substitutions, as opposed to single and multi-word substitutions for Szpektor *et al.*. Furthermore, our experiments are performed on 500 MB of text only, against 1 GB of news data for the DIRT system and the web for the TEASE algorithm. More data may help our recall, as well as bootstrapping over our best precision system.

We show a sample of our results in Table 6. The pairs with an asterisk were considered incorrect at human evaluation stage.

8 Conclusion

We have presented here a system for the extraction of word substitutions in the context of KB querying. We have shown that the output of a distributional similarity baseline can be improved by filtering it using the idea that two words in an entailment relation are substitutable in immediate surface contexts. We obtained a precision of 56% (52% using our most stringent evaluation) on a test set of 30 transitive verbs, and a yield of 3.4 relations per verb.

We also point out that relatively good precisions can be obtained on a parsed medium-sized corpus of 500 MB, although recall is certainly affected.

We note that our current implementation does not always satisfy the requirement for substitutability for generation tasks and point out that the system is therefore limited to our intended use, which involves search capabilities only.

We would like to concentrate in the future on providing a direction for the entailment pairs extracted by the system. We also hope that recall could possibly improve using a larger set of features in the pattern-based step (this is suggested also by Szpektor *et al.*, 2004), together with ap-

propriate bootstrapping.

Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC: EP/P502365/1). I would also like to thank my supervisor, Dr Ann Copestake, for her support throughout this project, as well as the anonymous reviewers who commented on this paper.

References

Enrique Alfonseca and Suresh Manandhar. 2002. *Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures*. In Proceedings of EKAW 2002, pp. 1–7, 2002.

Thorsten Brants and Alex Franz. 2006. *Web 1T 5-gram Version 1*. Linguistic Data Consortium, Philadelphia, 2006.

Edward Briscoe, John Carroll and Rebecca Watson. 2006. *The Second Release of the RASP System*. In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, Sydney, Australia, 2006.

Timothy Chklovski and Patrick Pantel. 2004. *VerbOcean: Mining The Web for Fine-Grained Semantic Verb Relations*. Proceedings of EMNLP-04, Barcelona, Spain, 2004.

Ann Copestake. 2004. Robust Minimal Recursion Semantics. `www.cl.cam.ac.uk/∼aac10/papers/rmrs draft.pdf`.

Ido Dagan and Oren Glickman. 2004. *Probabilistic Textual Entailment: Generic Applied Modelling of Language Variability*. Proceedings of The PASCAL Workshop on Learning Methods for Text Understanding and Mining, Grenoble, France, 2004.

Ido Dagan, Oren Glickman, Alfio Gliozzo, Efrat Marmorshtein and Carlo Strapparava. 2006. *Direct Word Sense Matching for Lexical Substitution*. Proceedings of COLING-ACL 2006, 17-21 Jul 2006, Sydney, Australia.

Maayan Geffet and Ido Dagan. 2004. *Feature Vector Quality and Distributional Similarity*. Proceedings Of the 20th International Conference on Computational Linguistics, 2004.

Maayan Geffet and Ido Dagan. 2005. *The Distributional Inclusion Hypothesises and Lexical Entailment*. In Proceedings Of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 107–114, 2005.

Mario Jarmasz and Stan Szpakowicz. 2003. *Roget's Thesaurus and Semantic Similarity*. In Proceedings of International Conference RANLP–03, pp. 212–219, 2003.

Zelig Harris. *Distributional Structure.* In Word, 10, No. 2–3, pp. 146–162, 1954.

Marti Hearst. 1992. *Automatic Acquisition of Hyponyms from Large Text Corpora*. Proceedings of COLING-92, pp.539–545, 1992.

Graeme Hirst and David St-Onge. 1998. *Lexical Chains As Representations of Context for the Detection and Correction of Malapropisms*. In 'WordNet', Ed. Christiane Fellbaum, Cambridge, MA: The MIT Press, 1998.

Dekang Lin. 2003. *An Information-Theoretic Definition of Similarity*. In Proceedings of the 15th International Conference on Machine Learning, pp. 296–304, 1998.

Dekang Lin, Shaojun Zhao, Lijuan Qin and Ming Zhou. 2003. *Identifying Synonyms among Distributionally Similar Words*. In Proceedings of IJCAI-03, Acapulco, Mexico, 2003.

Dekang Lin and Patrick Pantel. 2001. *DIRT – Discovery of Inference Rules from Text*. In Proceedings of ACM 2001, 2001.

George Miller and Walter Charles. 2001. *Contextual Correlates of Semantic Similarity*. In Language and Cognitive Processes, 6(1), pp. 1–28, 1991.

Shachar Mirkin, Ido Dagan and Maayan Geffet. 2004. *Integrating Pattern-Based and Distributional Similarity Methods for Lexical Entailment Acquisition*. In Proceedings of COLING/ACL, Sydney, Australia, pp.579–586, 2006.

Patrick Pantel and Deepak Ravichandran. 2004. *Automatically Labelling Semantic Classes*. In Proceedings of HLT/NAACL04, Boston, MA, pp 321328, 2004.

Deepak Ravichandran and Eduard Hovy. 2002. *Learning Surface Text Patterns for a Question Answering System*. Proceedings of ACL, 2002.

Philip Resnik. 1995. *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*. In Proceedings of IJCAI–95, 1995.

Idan Szpektor, Hristo Tanev, Ido Dagan and Bonaventura Coppola. 2004. *Scaling Web-Based Acquisition of Entailment Relations*. In Proceedings of EMNLP–2004, pp. 41–48, 2004.

Idan Szpektor, Eyal Shnarch and Ido Dagan. 2007. *Instance-Based Evaluation of Entailment Rule Acquisition*. In Proceedings of ACL–07, 2007.

Michael Strube and Simone Ponzetto. 2006. *WikiRelate! Computing Semantic Relatedness Using Wikipedia*. In Proceedings of AAAI–06, pp. 1219–1224, 2006.