What Substitutes Tell Us –
Analysis of an “All-Words” Lexical Substitution Corpus

Gerhard Kremer
Institute for Computational Linguistics
University of Heidelberg, Germany
kremer@cl.uni-heidelberg.de

Katrin Erk
Dept. of Linguistics
University of Texas, Austin, U.S.A.
katrin.erk@utexas.edu

Sebastian Padó
Institute for Natural Language Processing
University of Stuttgart, Germany
pado@ims.uni-stuttgart.de

Stefan Thater
Dept. of Computational Linguistics
Saarland University, Saarbrücken, Germany
stth@coli.uni-sb.de

Abstract

We present the first large-scale English “all-words lexical substitution” corpus. The size of the corpus provides a rich resource for investigations into word meaning. We investigate the nature of lexical substitute sets, comparing them to WordNet synsets. We find them to be consistent with, but more fine-grained than, synsets. We also identify significant differences to results for paraphrase ranking in context reported for the SEMEVAL lexical substitution data. This highlights the influence of corpus construction approaches on evaluation results.

1 Introduction

Many, if not most, words have multiple meanings; for example, the word “bank” has a financial and a geographical sense. One common approach to deal with this *lexical ambiguity* is supervised *word sense disambiguation*, or WSD (McCarthy, 2008; Navigli, 2009), which frames the task as a lemma-level classification problem, to be solved by training classifiers on samples of lemma instances that are labelled with their correct senses.

This approach has its problems, however. First, it assumes a complete and consistent set of labels. WordNet, used in the majority of studies, does cover several 10,000 lemmas, but has been criticised for both its coverage and granularity. Second, WSD requires annotation for each sense and lemma, leading to an “annotation bottleneck”. A number of technical solutions have been suggested regarding the second problem (Ando and Zhang, 2005; Navigli and Ponzetto, 2012), but not for the first.

In 2009, McCarthy and Navigli address both problems by proposing a fundamentally different approach, called *Lexical Substitution* (McCarthy and Navigli, 2009) which avoids capturing a word’s meaning by a single label. Instead, annotators are asked to list, for each instance of a word, one or more alternative words or phrases to be substituted for the target in this particular context. This setup provides a number of benefits over WSD. It allows characterising word meaning without using an ontology and can be obtained easily from native speakers through crowdsourcing. Work on modelling Lexical Substitution data has also assumed a different focus from WSD. It tends to see the prediction of substitutes along the lines of compositional lexical semantics, concentrating on explaining how word meaning is modulated in context (Mitchell and Lapata, 2010).

There are, however, important shortcomings of the work in the Lexical Substitution paradigm. All existing datasets (McCarthy and Navigli, 2009; Sinha and Mihalcea, 2014; Biemann, 2013; McCarthy et al., 2013) are either comparatively small, are “lexical sample” datasets, or both. “Lexical sample” datasets consist of sample sentences for each target word drawn from large corpora, with just one target word substituted in each sentence. In WSD, “lexical sample” datasets contrast with “all-words” annotation, in which all content words in a text are annotated for sense (Palmer et al., 2001).

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 540–549,
Gothenburg, Sweden, April 26-30 2014.

In this paper, we present the first large "all-words" Lexical Substitution dataset for English. It provides substitutions for more than 30,000 words of running text from two domains of MASC (Ide et al., 2008; Ide et al., 2010), a subset of the American National Corpus (`http://www.anc.org`) that is freely available and has (partial) manual annotation. The main advantage of the all-words setting is that it provides a realistic frequency distribution of target words and their senses. We use this to empirically investigate (a) the nature of lexical substitution and (b) the nature of the corpus, seen through the lens of word meaning in context.

2 Related Work

2.1 Lexical Substitution: Data

The original "English Lexical Substitution" dataset (McCarthy and Navigli, 2009) comprises 200 target content words (balanced numbers of nouns, verbs, adjectives and adverbs). Targets were explicitly selected to exhibit interesting ambiguities. For each target, 10 sentences were chosen (mostly at random, but in part by hand) from the English Internet Corpus (Sharoff, 2006) and presented to 5 annotators to collect substitutes. Its total size is 2,000 target instances. Sinha and Mihalcea (2014) produced a small pilot dataset (500 target instances) for all-words substitution, asking three annotators to substitute all content words in presented sentences.

Biemann (2013) first investigated the use of crowdsourcing, developing a three-task bootstrapping design to control for noise. His study covers over 50,000 instances, but these correspond only to 397 targets, all of which are high-frequency nouns. Biemann clusters the resulting substitutes into word senses. McCarthy et al. (2013) applied lexical substitution in a cross-lingual setting, annotating 130 of the original McCarthy and Navigli targets with Spanish substitutions (i. e., translations).

2.2 Lexical Substitution: Models

The LexSub task at SEMEVAL 2007 (McCarthy and Navigli, 2009) required systems to both determine substitution candidates and choose contextual substitutions in each case. Erk and Padó (2008) treated the gold substitution candidates as given and focused on the context-specific ranking of those candidates. In this form, the task has been addressed through three types of (mostly unsupervised) approaches. The first group computes a single type representation and modifies it according to sentence context (Erk and Padó, 2008; Thater et al., 2010; Thater et al., 2011; Van de Cruys et al., 2011). The second group of approaches clusters instance representations (Reisinger and Mooney, 2010; Dinu and Lapata, 2010; Erk and Padó, 2010; O'Séaghdha and Korhonen, 2011). The third option is to use a language model (Moon and Erk, 2013). Recently, supervised models have emerged (Biemann 2013; Szarvas et al., 2013a,b).

3 CoInCo – The MASC All-Words Lexical Substitution Corpus[1]

Compared to, e. g., WSD, there still is little gold-annotated data for lexical substitution. With the exception of the dataset created by Biemann (2013), all existing lexical substitution datasets are fairly small, covering at most several thousand instances and few targets which are manually selected. We aim to fill this gap, providing a dataset that mirrors the actual corpus distribution of targets in sentence context and is sufficiently large to enable a detailed, lexically specific analysis of substitution patterns.

3.1 Source Corpus Choice

For annotation, we chose a subset of the "Manually Annotated Sub-Corpus" MASC (Ide et al., 2008; Ide et al., 2010) which is "equally distributed across 19 genres, with manually produced or validated annotations for several layers of linguistic phenomena", created with the purpose of being "free of usage and redistribution restrictions". We chose this corpus because (a) our analyses can profit from the preexisting annotations and (b) we can release our annotations as part of MASC.

Since we could not annotate the complete MASC, we selected (complete) text documents from two prominent genres: *news* (18,942 tokens) and *fiction* (16,605 tokens). These two genres are both relevant for NLP and provide long, coherent documents that are appropriate for all-words annotation. We used the MASC part-of-speech annotation to identify all content words (verbs, nouns, adjectives, and adverbs), which resulted in a total of over 15,000 targets for annotation. This method differs from Navigli and McCarthy's (2009) in two crucial respects: we annotate all instances of each target, and include all targets regardless of frequency or level of lexical ambiguity. We believe that our corpus is considerably more representative of running text.

[1] Available as XML-formatted corpus "Concepts in Context" (CoInCo) from `http://goo.gl/5C0jBH`. Also scheduled for release as part of MASC.

3.2 Crowdsourcing

We used the Amazon Mechanical Turk (AMT) platform to obtain substitutes by crowdsourcing. Inter-annotator variability and quality issues due to non-expert annotators are well-known difficulties (see, e. g., Fossati et al. (2013)). Our design choices were shaped by "best practices in AMT", including Mason and Suri (2012) and Biemann (2013).

Defining HITs. An AMT task consists of Human Intelligence Tasks (HITs), each of which is supposed to represent a minimal, self-contained task. In our case, potential HITs were annotations of (all target words in) one sentence, or just one target word. The two main advantages of annotating a complete sentence at a time are (a) less overhead, because the sentence has only to be read once; (b) higher reliability, since all words within a sentence will be annotated by the same person.

Unfortunately, presenting individual sentences as HITs also means that all sentences pay the same amount irrespective of their length. Since long sentences require more effort, they are likely to receive less attention. We therefore decided to generally present two random target words per HIT, and one word in the case of "leftover" singleton targets.

In the HITs, AMT workers ("turkers") saw the highlighted target word in context. Since one sentence was often insufficient to understand the target fully, we also showed the preceding and the following sentence. The task description asked turkers to provide (preferably single-word) substitutes for the target that "would not change the meaning". They were explicitly allowed to use a "more general term" in case a substitute was hard to find (e. g., *dog* for the target *dachshund*, cf. basic level effects: Rosch et al. (1976)). Turkers were encouraged to produce as many replacements as possible (up to 5). If they could not find a substitute, they had to check one of the following radio buttons: "proper name", "part of a fixed expression", "no replacement possible", "other problem (with description)".

Improving Reliability. Another major problem is reliability. Ideally, the complete dataset should be annotated by the same group of annotators, but turkers tend to work only on a few HITs before switching to other AMT jobs. Following an idea of Biemann and Nygaard (2010), we introduced a two-tier system of jobs aimed at boosting turker loyalty. A tier of "open tasks" served to identify reliable turkers by manually checking their given substitutes for plausibility. Such turkers were then invited to the second, "closed task" tier, with a higher payment. In both tiers, bonus payments were offered to those completing full HIT sets.

For each target, we asked 6 turkers to provide substitutions. In total, 847 turkers participated successfully. In the open tasks, 839 turkers submitted 12,158 HITs (an average of 14.5 HITs). In the closed tasks, 25 turkers submitted 42,827 HITs (an average of 1,713 HITs), indicating the substantial success of our turker retention scheme.

Cost. In the open task, each HIT was paid for with $ 0.03, in the closed task the wage was $ 0.05 per HIT. The bonus payment for completing a HIT set amounted to $ 2 ($ 1) in the open (closed) tasks. The average cost for annotations was $ 0.22 for one target word instance and $ 0.02 for one substitute. The total cost with fees was ~$ 3,400.

3.3 COINCO: Corpus and Paraset Statistics

We POS-tagged and lemmatised targets and substitutes in sentence context with TreeTagger (Schmid, 1994). We manually lemmatised unknown words. Our annotated dataset comprises a total of 167,336 responses by turkers for 15,629 target instances in 2,474 sentences (7,117 nouns, 4,617 verbs, 2,470 adjectives, and 1,425 adverbs). As outlined above, targets are roughly balanced across the two genres (news: 8,030 instances in 984 sentences; fiction: 7,599 instances in 1,490 sentences). There are 3,874 unique target lemmas; 1,963 of these occur more than once. On this subset, there is a mean of 6.99 instances per target lemma. To our knowledge, our corpus is the largest lexical substitution dataset in terms of lemma coverage.

Each target instance is associated with a *paraset* (i. e., the set of substitutions or paraphrases produced for a target in its context) with an average size of 10.71. Turkers produced an average of 1.68 substitutions per target instance.[2] Despite our instructions to provide single-word substitutes, 11,337 substitutions contain more than one word.

3.4 Inter-Annotator Agreement

McCarthy and Navigli (2009) introduced two inter-annotator agreement (IAA) measures for their dataset. The first one is *pairwise agreement* (PA),

[2] Note that a small portion of the corpus was annotated by more than 6 annotators.

dataset	# targets	PA	mode-%	PA_m
MN09	1,703	27.7	73.9	50.7
SM13	550	15.5	N/A	N/A
COINCO (complete)	15,400	19.3	70.9	44.7
COINCO (subset)	2,828	24.6	76.4	50.9

Table 1: Pairwise turker agreement (mode-%: percentage of target instances with a mode)

measuring the overlap of produced substitutions:

$$\text{PA} = \sum_{t \in T} \sum_{\langle s_t, s'_t \rangle \in C_t} \frac{|s_t \cap s'_t|}{|s_t \cup s'_t|} \cdot \frac{1}{|C_t| \cdot |T|}$$

where t is a target in our target set T, s_t is the paraset provided by one turker for t, and C_t is the set comprising all pairs of turker-specific parasets for t. Only targets with non-empty parasets (i. e., not marked by turkers as a problematic target) from at least two turkers are included. The second one is *mode agreement* (PA_m), the agreement of annotators' parasets with the *mode* (the unique most frequent substitute) for all targets where one exists:

$$\text{PA}_\text{m} = \sum_{t \in T_m} \sum_{s_t \in S_t} [m \in s_t] \cdot \frac{1}{|s_t| \cdot |T_m|}$$

where T_m is the set of all targets with some mode m and S_t is the set of all parasets for target t. The Iverson bracket notation $[m \in s_t]$ denotes 1 if mode m is included in s_t (otherwise 0).

Table 1 compares our dataset to the results by McCarthy and Navigli (2009, MN09) and Sinha and Mihalcea (2014, SM13). The scores for our complete dataset (row 3) are lower than McCarthy and Navigli's both for PA (−8 %) and PA_m (−6 %), but higher than Sinha and Mihalcea's, who also note the apparent drop in agreement.[3]

We believe that this is a result of differences in the setup rather than an indicator of low quality: Note that PA will tend to decrease both in the face of more annotators and of more substitutes. Both of these factors are present in our setup. To test this interpretation, we extracted a subset of our data that is comparable to McCarthy and Navigli's regarding these factors. It comprises all target instances where (a) exactly 6 turkers gave responses (9,521 targets), and (b) every turker produced between one and three substitutes (5,734 targets). The results for this subset (row 4) are much more similar to those of McCarthy and Navigli: the pairwise agreement differs only by 3 %, and the mode agreement is almost identical. We take these figures as indication that crowdsourcing can serve as a sufficiently reliable way to create substitution data; note that Sinha and Mihalcea's annotation was carried out "traditionally" by three annotators.

relation	all	verb	noun	adj	adv
syn	9.4	12.5	7.7	8.0	10.4
direct-hyper	6.6	9.3	7.6	N/A	N/A
direct-hypo	7.5	11.6	8.0	N/A	N/A
trans-hyper	3.2	2.8	4.7	N/A	N/A
trans-hypo	3.0	3.7	3.8	N/A	N/A
wn-other	68.9	60.7	66.5	88.5	85.4
not-in-wn	2.1	0.9	2.2	3.4	4.2

Table 2: Target–substitute relations in percentages, overall (*all*) and by POS. Note: WordNet contains no hypo-/hypernyms for adjectives and adverbs.

Investigating IAA numbers by target POS and by genre, we found only small differences (≤ 2.6 %) among the various subsets, and no patterns.

4 Characterising Lexical Substitutions

This section examines the collected lexical substitutions, both quantitatively and qualitatively. We explore three questions: (a) What lexical relations hold between targets and their substitutes? (b) Do parasets resemble word senses? (c) How similar are the parasets that correspond to the same word sense of a target? These questions have not been addressed before, and we would argue that they could not be addressed before, because previous corpora were either too small or were sampled in a way that was not conducive to this analysis.

We use WordNet (Fellbaum, 1998), release 3.1, as a source for both lexical relations and word senses. WordNet is the de facto standard in NLP and is used for both WSD and broader investigations of word meaning (Navigli and Ponzetto, 2012; Erk and McCarthy, 2009). Multi-word substitutes are excluded from all analyses.[4]

4.1 Relating Targets and Substitutes

We first look at the most canonical lexical relations between a target and its substitutes. Table 2 lists the percentage of substitutes that are synonyms (*syn*), direct/transitive (*direct-*/*trans-*) hypernyms (*hyper*)

[3] Please see McCarthy and Navigli (2009) for a possible explanation of the generally low IAA numbers in this field.

[4] All automatic lexical substitution approaches, including Section 5, omit multi-word expressions. Also, they can be expected to have WordNet coverage and normalisation issues, which would constitute a source of noise for this analysis.

sentence	substitutes
Now, how can I help the elegantly mannered friend of my Nepthys and his surprising young charge ?	dependent, person, *task*, *lass*, *protégé*, *effort*, *companion*
The distinctive whuffle of pleasure rippled through the betas on the bridge, and Rakal let loose a small growl, as if to caution his charges against false hope.	dependent, command, accusation, *private*, *companion*, *follower*, *subordinate*, *prisoner*, *teammate*, *ward*, *junior*, *underling*, *enemy*, *group*, *crew*, *squad*, *troop*, *team*, *kid*

Table 3: Context effects below the sense level: target noun "charge" (wn-other shown in italics)

and hyponyms (*hypo*) of the target. If a substitute had multiple relations to the target, the shortest path from any of its senses to any sense of the target was chosen. The table also lists the percentage of substitutes that are elsewhere in WordNet but not related to the target (*wn-other*) and substitutes that are not covered by WordNet (*not-in-wn*).

We make three main observations. First, WordNet shows very high coverage throughout – there are very few *not-in-wn* substitutes. Second, the percentages of synonyms, hypernyms and hyponyms are relatively similar (even though the annotation guidelines encouraged the annotation of hyponyms over hypernyms), but relatively small. Finally, and most surprisingly, the vast majority of substitutes across all parts of speech are *wn-other*.

A full analysis of *wn-other* is beyond the current paper. But a manual analysis of *wn-other* substitutes for 10 lemmas[5] showed that most of them were *context-specific* substitutes that can differ even when the sense of the target is the same. This is illustrated in Table 3, which features two occurrences of the noun "charge" in the sense of "person committed to your care". But because of the sentence context, the first occurrence got substitutes like "protégé", while the second one was paraphrased by words like "underling". We also see evidence of annotator error (e. g., "command" and "accusation" in the second sentence).[6] Discounting such instances still leaves a prominent role for correct *wn-other* cases.

But are these indeed contextual modulation effects below the sense level, or are parasets fundamentally different from word senses? We perform two quantitative analyses to explore this question.

4.2 Comparing Parasets to Synsets

To what extent do parasets follow the boundaries of WordNet senses? To address this question, we establish a mapping between parasets and synsets. Since gold standard word senses in MASC are limited to high-frequency lemmas and cover only a small part of our data, we create a heuristic mapping that assigns each paraset to that synset of its target with which it has the largest intersection. We use *extended WordNet synsets* that include direct hypo- and hypernyms to achieve better matches with parasets. We call a paraset *uniquely mappable* if it has a unique best WordNet match, and *mappable* if one or more best matches exist. Table 4 shows that most parasets are mappable for nouns and verbs, but not for adjectives or adverbs.

paraset–sense mapping class	verb	noun	adj	adv
mappable	90.3	73.5	33.0	49.6
uniquely mappable	63.1	57.5	24.3	41.3

Table 4: Ratios of (uniquely) mappable parasets

We now focus on mappable parasets for nouns and verbs. To ensure that this does not lead to a confounding bias, we performed a small manual study on the 10 noun and verb targets mentioned above (247 parasets). We found 25 non-mappable parasets, which were due to several roughly equally important reasons: gaps in WordNet, multi-word expressions, metaphor, problems of sense granularity, and annotator error. We also found 66 parasets with multiple best matches. The two dominant sources were target occurrences that evoked more than one sense and WordNet synset pairs with very close meanings. We conclude that excluding non-mappable parasets does not invalidate our analysis.

To test whether parasets tend to map to a single synset, we use a cluster purity test that compares a set of clusters C to a set of gold standard classes C'. Purity measures the accuracy of each cluster with respect to its best matching gold class:

$$purity(C, C') = \frac{1}{N} \sum_{k=1}^{K} \max_{k'} |C_k \cap C'_{k'}|$$

where N is the total number of data points, K is the

[5] We used the nouns *business, charge, place, way* and the verbs *call, feel, keep, leave, show, stand.*

[6] A manual analysis of the same 10 lemmas showed only 38 out of 1,398 (0.027) of the substitutes to be erroneous.

measure	verbs	nouns
cluster purity (%)	75.1	81.2
common core size within sense	1.84	2.21
common core size across senses	0.39	0.41
paraset size	6.89	6.29

Table 5: Comparing uniquely mappable parasets to senses: overlap with best WordNet match as cluster purity (top), and intersection size of parasets with and without the same WordNet match (bottom)

set	elements
synset \ core	feel, perceive, comprehend
synset ∩ core	sense
core \ synset	notice
non-core substitutes	detect, recall, perceive, experience, note, realize, discern

Table 6: Target feel.v.03: synset and common core

number of clusters, and $C'_{k'}$ is the gold class that has the largest overlap with cluster C_k. In our case, C is the set of mappable parasets[7], C' the set of extended WordNet synsets, and we only consider substitutes that occur in one of the target's extended synsets (these are the data points). This makes the current analysis complementary to the relational analysis in Table 2.[8]

The result, listed in the first row of Table 5, shows that parasets for both verbs and nouns have a high purity, that is, substitutes tend to focus on a single sense. This can be interpreted as saying that annotators tend to agree on the general sense of a target. Roughly 20–25 % of substitutes, however, tend to stem from a synset of the target that is not the best WordNet match. This result comes with the caveat that it only applies to substitutes that are synonyms or direct hypo- and hypernyms of the target. So in the next section, we perform an analysis that also includes *wn-other* substitutes.

4.3 Similarity Between Same-Sense Parasets

We now use the WordNet mappings from the previous section to ask how (dis-)similar parasets are that represent the same word sense. We also try to identify the major sources for dissimilarity.

We quantify paraset similarity as the *common core*, that is, the intersection of *all* parasets for the same target that map onto the same extended WordNet synset. Surprisingly, the common core is mostly non-empty (in 85.6 % of all cases), and contains on average around two elements, as the second row in Table 5 shows. For this analysis, we only use uniquely mappable parasets. In relation to the average paraset size (see row 4), this means that one quarter to one third of the substitutes are shared among all instances of the same target–sense combination. In contrast, the common core for all parasets of targets that map onto two or more synsets contains only around 0.4 substitutes (see row 3) – that is, it is empty more often than not.

At the same time, if about one quarter to one third of the substitutes are shared, this means that there are more non-shared than shared substitutes even for same-sense parasets. Some of these cases result from small samples: Even 6 annotators cannot always exhaust all possible substitutes. For example, the phrase "I'm starting to see more *business* transactions" occurs twice in the corpus. The two parasets for "business" share the same best WordNet sense match, but they have only 3 shared and 7 non-shared substitutes. This is even though the substitutes are all valid and apply to both instances. Other cases are instances of the context sensitivity of the Lexical Substitution task as discussed above. Table 6 illustrates on an example how the common core of a target sense relates to the corresponding synset; note the many context-specific substitutes outside the common core.

5 Ranking Paraphrases

While there are several studies on modelling lexical substitutes, almost all reported results use McCarthy and Navigli's SEMEVAL 2007 dataset. We now compare the results of three recent computational models on COINCO (our work) and on the SEMEVAL 2007 dataset to highlight similarities and differences between the two datasets.

Models. We consider the paraphrase ranking models of Erk and Padó (2008, EP08), Thater et al. (2010, TFP10) and Thater et al. (2011, TFP11). These models have been analysed by Dinu et al. (2012) as instances of the same general framework and have been shown to deliver state-of-the-art performance on the SEMEVAL 2007 dataset, with best results for Thater et al. (2011).

The three models share the idea to represent the meaning of a target word in a specific context by

[7] For non-uniquely mappable parasets, the purity is the same for all best-matching synsets.

[8] Including *wn-other* substitutes would obscure whether low purity means substitutes from a mixture of senses (which we are currently interested in) or simply a large number of *wn-other* substitutes (which we have explored above).

corpus		syntactically structured			syntactically filtered		bag of words		random
		Tfp11	Tfp10	Ep08	Tfp11/Ep08	Tfp10	Tfp11/Ep08	Tfp10	
CoInCo	context	**47.8**	46.0	47.4	47.4	41.9	46.2	40.8	33.0
	baseline	46.2	44.6	46.2	45.8	38.8	44.7	37.5	
SemEval 2007	context	**52.5**	48.6	49.4	50.1	44.7	48.0	42.6	30.0
	baseline	43.7	42.7	43.7	44.4	38.0	42.7	35.8	
CoInCo Subset	context	**40.3**	37.7	39.0	39.2	34.1	37.7	32.5	23.7
	baseline	36.7	35.7	36.7	36.4	30.6	35.4	28.0	

Table 7: Corpus comparison in terms of paraphrase ranking quality (Gap percentage). SemEval results from Thater et al. (2011). “Context”: full models, “baseline”: uncontextualised target-substitute similarity.

modifying the target’s basic meaning vector with information from the vectors of the words in the target’s direct syntactic context. For instance, the vector of “coach” in the phrase “the coach derailed” is obtained by modifying the basic vector representation of “coach” through the vector of “derail”, so that the resulting contextualised vector reflects the train car sense of “coach”.

We replicate the setup of Thater et al. (2011) to make our numbers directly comparable. We consider three versions of each model: (a) *syntactically structured* models use vectors which record co-occurrences based on dependency triples, explicitly recording syntactic role information within the vectors; (b) *syntactically filtered* models also use dependency-based co-occurrence information, but the syntactic role is not explicitly represented in the vector representations; (c) *bag-of-words* models use a window of ± 5 words. All co-occurrence counts are extracted from the English Gigaword corpus (`http://catalog.ldc.upenn.edu/LDC2003T05`), analysed with Stanford dependencies (de Marneffe et al., 2006).

We apply the models to our dataset as follows: We first collect all substitutes for all occurrences of a target word in the corpus. The task of our models for each target instance is then to rank the candidates so that the actual substitutes are ranked higher than the rest. We rank candidates according to the cosine similarity between the contextualised vector of the target and the vectors of the candidates. Like most previous approaches, we compare the resulting ranked list with the gold standard annotation (the paraset of the target instance), using generalised average precision (Kishida, 2005, Gap), and using substitution frequency as weights. Gap scores range between 0 and 1; a score of 1 indicates a perfect ranking in which all correct substitutes precede all incorrect ones, and correct high-weight substitutes precede low-weight substitutes.

Results. The upper part of Table 7 shows results for our CoInCo corpus and the previous standard dataset, SemEval 2007. “Context” refers to the full models, and “baseline” to global, context-unaware ranking based on the semantic similarity between target and substitute. Baselines are model-specific since they re-use the models’ vector representations. Note that Ep08 and Tfp11 are identical unless syntactically structured vectors are used, and their baselines are identical.

The behaviour of the baselines on the two corpora is quite similar: random baselines have Gaps around 0.3, and uncontextualised baselines have Gaps between 0.35 and 0.46. The order of the models is also highly parallel: the syntactically structured Tfp11 is the best model, followed by its syntactically filtered version and syntactically structured Ep08. All differences between these models are significant ($p < 0.01$) for both corpora, as computed with bootstrap resampling (Efron and Tibshirani, 1993). That is, the model ranking on SemEval is replicated on CoInCo.

There are also substantial differences between the two corpora, though. Most notably, all models perform substantially worse on CoInCo. This is true in absolute terms (we observe a loss of 2–5 % Gap) but even more dramatic expressed as the gain over the uninformed baselines (almost 9 % for Tfp11 on SemEval but only 1.2 % on CoInCo). All differences between CoInCo and SemEval are again significant ($p < 0.01$).

We see three major possible reasons for these differences: variations in (a) the annotation setup (crowdsourcing, multiple substitutes); (b) the sense distribution; (c) frequency and Pos distributions between the two corpora. We focus on (c) since it can be manipulated most easily. SemEval contains exactly 10 instances for all targets, while CoInCo reflects the Zipf distribution of “natural” corpora, with many targets occurring only once. Such corpora are easier to model in terms of absolute performance, because the paraphrase lists for rare targets contain less false positives for each instance. For hapax legomena, the set of substitution candidates is identical to the gold standard, and the only way to receive a GAP score lower than 1 for such targets is to rank low-weight substitutes ahead of high-weight substitutes. Not surprisingly, the mean GAP score of the syntactically structured TFP11 for hapax legomena is 0.863. At the same time, such corpora make it harder for full models to outperform uncontextualised baselines; the best model (TFP11) only outperforms the baseline by 1.6 %.

To neutralise this structural bias, we created "SEMEVAL-like" subsets of COINCO (collectively referred to as the COINCO Subset) by extracting all COINCO targets with at least 10 instances (141 nouns, 101 verbs, 50 adjectives, 36 adverbs) and building 5 random samples by drawing 10 instances for each target. These samples match SEMEVAL in the frequency distribution of its targets. To account for the unequal distribution of POS in the samples, we compute GAP scores for each POS separately and calculate these GAP scores' average.

The results for the various models on the COINCO Subset in the bottom part of Table 7 show that the differences between COINCO and SEMEVAL are not primarily due to the differences in target frequencies and POS distribution – the COINCO Subset is actually more different to SEMEVAL than the complete COINCO. Strikingly, the COINCO Subset is very difficult, with a random baseline of 24 % and model performances below 37 % (baselines) and up to 40 % (full models), which indicates that the set of substitutes in COINCO is more varied than in SEMEVAL as an effect of the annotation setup. Encouragingly, the margin between full models and baselines is larger than on the complete COINCO and generally amounts to 2–4 % (3.6 % for TFP11). That is, the full models are more useful on the COINCO corpus than they appeared at first glance; however, their effect still remains much smaller than on SEMEVAL.

6 Conclusion

This paper describes COINCO, the first large-scale "all-words" lexical substitution corpus for English. It was constructed through crowdsourcing on the basis of MASC, a corpus of American English.

The corpus has two major advantages over previous lexical substitution corpora. First, it covers contiguous documents rather than selected instances. We believe that analyses on our corpus generalise better to the application domain of lexical substitution models, namely random unseen text. In fact, we find substantial differences between the performances of paraphrase ranking models for COINCO and the original SEMEVAL 2007 LexSub dataset: the margin of informed methods over the baselines are much smaller, even when controlling for target frequencies and POS distribution. We attribute this divergence at least in part to the partially manual selection strategy of SEMEVAL 2007 (cf. Section 2.1) which favours a more uniform distribution across senses, while our whole-document annotation faces the "natural" distribution skewed towards predominant senses. This favours the non-contextualised baseline models, consistent with our observations. At the very least, our findings demonstrate the sensitivity of evaluation results on corpus properties.

The second benefit of our corpus is that its size enables more detailed analyses of lexical substitution data than previously possible. We are able to investigate the nature of the paraset, i. e., the set of lexical substitutes given for one target instance, finding that lexical substitution sets correspond fairly well to WordNet sense distinctions (parasets for the same synset show high similarity, while those for different senses do not). In addition, however, we observe a striking degree of context-dependent variation below the sense level: the majority of lexical substitutions picks up fine-grained, situation-specific meaning components that do not qualify as sense distinctions in WordNet.

Avenues for future work include a more detailed analysis of the substitution data to uncover genre- and domain-specific patterns and the development of lexical substitution models that take advantage of the all-words substitutes for global optimisation.

Acknowledgements

We are grateful to Jan Pawellek for implementing the AMT task, extracting MASC data, and preparing HITS. Furthermore, we thank Georgiana Dinu for her support with the word meaning models.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Chris Biemann and Valerie Nygaard. 2010. Crowdsourcing WordNet. In *Proceedings of the 5th Global WordNet conference*, Mumbai, India.

Chris Biemann. 2013. Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*, 47(1):97–122.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, pages 449–454, Genoa, Italy.

Georgiana Dinu and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of EMNLP*, pages 1162–1172, Cambridge, MA.

Georgiana Dinu, Stefan Thater, and Sören Laue. 2012. A comparison of models of word meaning in context. In *Proceedings of NAACL*, pages 611–615, Montréal, Canada.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proceedings of EMNLP*, pages 440–449, Singapore.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of EMNLP*, pages 897–906, Honolulu, HI.

Katrin Erk and Sebastian Padó. 2010. Exemplar-based models for word meaning in context. In *Proceedings of ACL*, pages 92–97, Uppsala, Sweden.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge, MA.

Marco Fossati, Claudio Giuliano, and Sara Tonelli. 2013. Outsourcing FrameNet to the crowd. In *Proceedings of ACL*, pages 742–747, Sofia, Bulgaria.

Nancy Ide, Collin F. Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. MASC: The manually annotated sub-corpus of American English. In *Proceedings of LREC*, pages 2455–2461, Marrakech, Morocco.

Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau. 2010. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of ACL*, pages 68–73, Uppsala, Sweden.

Kazuaki Kishida. 2005. Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments. Technical Report NII-2005-014E, Japanese National Institute of Informatics.

Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1):1–23.

Diana McCarthy and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

Diana McCarthy, Ravi Sinha, and Rada Mihalcea. 2013. The cross-lingual lexical substitution task. *Language Resources and Evaluation*, 47(3):607–638.

Diana McCarthy. 2008. Word sense disambiguation. In *Linguistics and Language Compass*. Blackwell.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Taesun Moon and Katrin Erk. 2013. An inference-based model of word meaning in context as a paraphrase distribution. *ACM Transactions on Intelligent Systems and Technology*, 4(3).

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41:1–69.

Diarmuid O'Séaghdha and Anna Korhonen. 2011. Probabilistic models of similarity in syntactic context. In *Proceedings of EMNLP*, pages 1047–1057, Edinburgh, UK.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *Proceedings of the SENSEVAL-2 workshop*, pages 21–24, Toulouse, France.

Joseph Reisinger and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Proceeding of NAACL*, pages 109–117, Los Angeles, CA.

Eleanor Rosch, Carolyn B. Mervis, Wayne D. Gray, David M. Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of NEMLAP*, pages 44–49, Manchester, UK.

Serge Sharoff. 2006. Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462.

Ravi Sinha and Rada Mihalcea. 2014. Explorations in lexical sample and all-words lexical substitution. *Natural Language Engineering*, 20(1):99–129.

György Szarvas, Chris Biemann, and Iryna Gurevych. 2013a. Supervised all-words lexical substitution using delexicalized features. In *Proceedings of NAACL-HLT*, pages 1131–1141, Atlanta, GA.

György Szarvas, Róbert Busa-Fekete, and Eyke Hüllermeier. 2013b. Learning to rank lexical substitutions. In *Proceedings of EMLNP*, pages 1926–1932, Seattle, WA.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. In *Proceedings of ACL*, pages 948–957, Uppsala, Sweden.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2011. Word meaning in context: A simple and effective vector model. In *Proceedings of IJCNLP*, pages 1134–1143, Chiang Mai, Thailand.

Tim Van de Cruys, Thierry Poibeau, and Anna Korhonen. 2011. Latent vector weighting for word meaning in context. In *Proceedings of EMNLP*, pages 1012–1022, Edinburgh, Scotland.