Unsupervised Ranked Cross-Lingual Lexical Substitution for Low-Resource Languages

Stefan Ecker, Andrea Horbach, Stefan Thater
Saarland University
Department of Computational Linguistics
Saarbrücken, Germany
{stefane, andrea, stth}@coli.uni-saarland.de

Abstract

We propose an unsupervised system for a variant of cross-lingual lexical substitution (CLLS) to be used in a reading scenario in computer-assisted language learning (CALL), in which single-word translations provided by a dictionary are ranked according to their appropriateness in context. In contrast to most alternative systems, ours does not rely on either parallel corpora or machine translation systems, making it suitable for low-resource languages as the language to be learned. This is achieved by a graph-based scoring mechanism which can deal with ambiguous translations of context words provided by a dictionary. Due to this decoupling from the source language, we need monolingual corpus resources only for the target language, i.e. the language of the translation candidates. We evaluate our approach for the language pair Norwegian Nynorsk–English on an exploratory manually annotated gold standard and report promising results. When running our system on the original SemEval CLLS task, we rank 6th out of 18 (including 2 baselines and our 2 system variants) in the *best* evaluation.

Keywords: cross-lingual lexical substitution, low-resource languages, computer-assisted language learning

1. Introduction

Over the last few years, online learning platforms for foreign languages have emerged to complement and sometimes even replace classroom-based courses. These platforms are particularly important for those languages which are not offered at local language schools.

One technique suitable for self-study CALL applications is *Intensive Reading*, where you try to gain a deep understanding of a small amount of text with the help of a dictionary. Yet dictionary usage itself poses a challenge to (at least beginning) language learners: if a word has several translations with different meanings, how does one know which is the right one in a particular context? A desirable feature of a CALL application for Intensive Reading would thus be a ranking mechanism that ranks the translations of a certain word depending on its context.

In the general case, one way to address this problem is to make use of the advances in cross-lingual lexical substitution (Sinha et al., 2009): When given a word with its context in language A, the task is to substitute the word with one or more words from language B that are good translations in that particular context. This task is rather resource-intensive: most system variants that took part in the SemEval-2010 Cross-Lingual Lexical Substitution Task use some sort of parallel corpora or a machine translation system (McCarthy et al., 2013). For low-resource languages, aligned parallel corpora or off-the-shelf machine translation programs are often not of comparable size and quality to the language pair Spanish–English used in that task, if such resources are available at all.

A common choice for the parallel corpus used in the CLLS systems is the sentence-aligned Europarl corpus for the language pair English–Spanish. Releases of comparable size are available for 9 additional languages (Koehn, 2005). Incidentally, most of these languages are already commonly offered in language schools and online learning platforms. The releases for the languages of the newer EU members are naturally much smaller in size. For machine translation, only three (English, French, Spanish) out of 30 European languages are reported to have better than fragmentary support (Ananiadou et al., 2012). Of course, many more languages in the world are of potential interest to learners. For many of those languages, the availability and quality of language resources relevant to this task are likely to be worse.

We propose a system which tackles both the problem of limited availability of parallel corpora and quality of dictionary output, by disambiguating all possible translations for a word given all possible translations for surrounding context words. For the translations, we rely on a dictionary, which is needed for the learning platform anyway. To do so, we use the *PageRank* algorithm to rank potential translations of a given word according to how well they are connected to possible translations of words from the context of the given word in an undirected graph. The main intuition is that good translations have stronger links into the context than translations that do not fit the context.

Take for example the word *disk* in American English. Besides the general meaning, it can also refer to more specific objects, such as a *diskette* or a *spinal disk*, which might require a translation unrelated to the general meaning of *disk* in other languages. If this word is embedded in a context featuring *computer*, *insert* and *drive*, then the translation of *disk* meaning *diskette* will co-occur with some translations of each context word in a corpus in the target language. The translation of *spinal disk* might also co-occur with some of these translations, but probably to a lesser degree, i.e. it will have fewer edges in the disambiguation graph, and the edges will have smaller weights.

Note that our task is slightly different from the original CLLS task: While in CLLS, systems have to propose good translations, in our setting the task is to rank a pool of given translations (those proposed by a potentially noisy dictionary).

As an example for a low-resource language, we chose Norwegian Nynorsk. Nynorsk is one of two official standard variants of Norwegian, but is used by only about 10% of the population in Norway, or around 500,000 people. For the evaluation of our system on the language pair Nynorsk-English, we introduce a new gold standard corpus. Our results show an improvement of 26.5% compared to a random baseline and 14.2% compared to an informed baseline for finding a single good translation.

Additionally, we also evaluate our method on the SemEval-2010 data, showing that we can reach results that are competitive with systems that make use of a much richer basis of resources, such as parallel corpora.

After this introduction, we will, in the remainder of the paper, present an overview of related work in the field of lexical substitution in Section 2. We next describe our corpus collection and annotation in Section 3, and present the general idea of our system in Section 4. We present and discuss our results in Section 5, before evaluating our system on the original SemEval-2010 task in Section 6 and concluding in Section 7.

2. Related Work

Most of the proposed systems for cross-lingual lexical substitution were presented for the SemEval-2010 Cross-Lingual Lexical Substitution Task (Mihalcea et al., 2010). Out of the 15 system variants that took part in the task, 11 use parallel corpora in of way or the other. A parallel corpus is also used in a more recent contribution by Apidianaki (2011).

Systems for SemEval-2010 not using parallel corpora do lexical substitution on either the source language side (TYO, SWAT-E) or in the target language (SWAT-S, UBA-T) assuming an unambiguously translated context sentence (Mihalcea et al., 2010; Wicentowski et al., 2010; Basile and Semeraro, 2010). Also, two more recent contributions (Sinha, 2013; Zahran et al., 2015) use a machine translation system to translate the context sentence before applying their method of contextual disambiguation between substitution candidates.

Systems using parallel corpora are not suitable for most low-resource languages, as the availability of such resources cannot be assumed. For our language pair Nynorsk–English we do not have access to a suitable parallel corpus. Systems using off-the-shelf machine translation systems could be considered when the language pair in question is available. However, not all available language pairs work as well as English–Spanish. While Norwegian–English is available on Google Translate, the service does not distinguish between Norwegian Nynorsk and Norwegian Bokmål. A manual tryout revealed that it does not work so well with Nynorsk; many words are left untranslated. Consequently, none of the described systems is directly applicable to our task.

Cross-Lingual Lexical Substitution could also be viewed as word sense disambiguation with translations as the sense descriptors. FCC-LS (Vilarino et al., 2010), for example, makes use of cross-lingual WSD as a first step for lexical substitution, Apidianaki (2011) does word sense induction on a parallel corpus. Mihalcea (2005) proposes an unsupervised graph-based algorithm for WSD, which we hypothesize would also work with simple lemmas instead of a predefined sense inventory, and would therefore be suitable for both monolingual and cross-lingual lexical substitution.

	Noun	**Verb**	**Adj**	**All**
Lemmas	8	7	5	20
Contexts	40	35	25	100
Raters/context	2.60	2.40	2.3	2.47
Translations/lemma	6.25	8.71	12.2	8.60

Table 1: Corpus overview

3. Data and Annotation

Since we want to test the performance of our system specifically for low-resource languages, no suitable evaluation data could be obtained from any standard tasks. Therefore, we constructed a small evaluation corpus containing 20 ambiguous lemmas selected from a set of high- and medium-frequency words in Norwegian Nynorsk, in the style of the English lexical substitution task (McCarthy and Navigli, 2009). For each lemma, five different context snippets were selected in part from the Norsk Ordbok's Nynorsk corpus (Norsk Ordbok 2014, 2012) and in part from the web, in both cases mainly extracts of online newspaper articles. The only constraint for the selection was that at least two different meanings of the lemma had to be contained among the five text snippets, so that non-synonymous translations provided by our dictionary (see Section 3.1) would also get evaluated. In sum, the evaluation corpus contains 100 test instances (i.e. lemma-context pairs), in which there are 40 nouns, 35 verbs and 25 adjectives. For each lemma, all dictionary translations into English were collected, leading to 172 lemma-translation pairs and 860 individual triples of lemma, translation and text context. A breakdown of the corpus statistics by word-class is shown in Table 1.

In an online crowd-sourcing annotation setup, we addressed both intermediate to advanced learners of Norwegian and native Scandinavian speakers for their rating of the quality of translations. All annotators indicated a good to excellent command of English. The continental Scandinavian languages are part of the same dialect continuum and mutually comprehensible at least in their written form (Maurud, 1976), so we also consider non-Norwegian Scandinavians to be good annotators. Annotators were shown a Nynorsk word in its lemmatized form together with the part-of-speech tag, the corresponding text snippet containing the word in its full form, and all English translations of the lemma, and were asked to rate each translation with one of the following labels: “good”, “acceptable” or “wrong”. Annotators could also indicate whether they did not understand the suggested translation and were asked not to look up unknown words. They were also able to skip questions if they did not understand the text in Nynorsk. We were mainly interested in how well the provided translations facilitate comprehension rather than their accuracy in an actual translation setting, so we suggested imagin-

	dune	blanket	eiderdown
Annotator 1	wrong	good	good
(Scandinavian)	(0)	(1)	(1)
Annotator 2	wrong	good	acceptable
(Scandinavian)	(0)	(1)	(0.5)
Annotator 3	wrong	acceptable	acceptable
(non-Scandinavian)	(0)	(0.5)	(0.5)
gold-scan rating	0	1.0	0.75
gold-all rating	0	0.9	0.70

Table 2: Example of gold standard construction from ratings.

ing a language-learning scenario when in doubt. Using this setup, we collected 247 annotations for individual Nynorsk words in context (i.e. 2.47 annotators responded per test instance) corresponding to 1927 ratings (2100 including unknown translations) of word-translation pairs from 7 different annotators (2 Norwegians, 2 other Scandinavians, 3 non-Scandinavians).

For the construction of gold standard ratings, we assigned confidence values per label: $good = 1.0$, $acceptable = 0.5$ and $wrong = 0.0$. These values were averaged per word-translation pair over all relevant annotators to obtain the final confidence score for a translation. From this data, we construct two different gold standards: The *gold-scan* data contains only ratings by Scandinavian native speakers in order to ensure that the annotators had a near perfect understanding of the Nynorsk text snippet. This setup got us at least one annotation for each item and at least a double annotation for 30%. The *gold-all* data contains all ratings, including those by language learners in order to consider the learner's view as well, especially with regard to the preference and comprehension of near-synonyms in English. In *gold-all*, the ratings of Scandinavians are given twice the weight of learners in order to resolve tie situations in favor of native speaker intuition. Including learners' ratings raises the proportion of double annotations (or more) to 85%.

As a simplified example, in the context *Mannen tidleg i 20-åra hadde pakka seg godt inn og lagt seg til rette med* ***dyne*** *og pute på sofaen til Eli*[1], the following translations for the noun *dyne* should be rated: *dune*, *blanket*, *eiderdown*. Table 2 shows an example of gold standard ratings obtained by taking the weighted average of annotator ratings using our two different strategies.

Since each annotator rated only a random sample of the corpus and could manually skip questions, we report the inter-rater agreement on a 20% subset of the evaluation corpus containing all lemmas once. This subset has been selected such that it contains the largest number of items that all have been annotated by the same set of four annotators (1 Norwegian, 2 other Scandinavians, 1 learner). The average pair-wise Cohen's weighted kappa for those three Scandinavians is 0.41; including the learner, it is 0.43. Thus there is moderate agreement between the annotators according to Landis and Koch (1977). As weights we used the numerical difference of the ratings. A breakdown by word classes and individual pairings is presented in Table 3. Nouns show better agreement than verbs and adjectives, most prominently for the pairing of the Norwegian and the Dane, which is the only pairing that shows substantial agreement at least for one word class.

Annotators	Noun	Verb	Adjective	All
Norwegian–Dane	0.74	0.39	0.32	0.49
Norwegian–Swede	0.58	0.30	0.35	0.44
Norwegian–Italian	0.52	0.45	0.59	0.53
Dane–Swede	0.49	0.25	0.22	0.32
Italian–Swede	0.44	0.42	0.21	0.38
Italian–Dane	0.43	0.54	0.28	0.42

Table 3: Inter-rater agreement for selected annotators using Cohen's weighted κ

3.1. Dictionary Statistics

Since we did not have access to a dictionary for Nynorsk–English with reasonable coverage, we had to induce one using a semi-automatic pipeline. This induced dictionary is used both for the gold standard and our system. We translate the lemmas first between the two Norwegian variants Nynorsk and Bokmål using the morphologically tagged dictionary that comes with the rule-based machine translation system Apertium (Unhammer and Trosterud, 2009), and then use a combination of glosbe.com and translations extracted from BabelNet 2.5 (Navigli and Ponzetto, 2012) for the step from Bokmål to English. For our dataset, the dictionary suggested on average 8.6 translations per lemma, 6.25 for nouns, 8.71 for verbs and 12.2 for adjectives.

An induced dictionary from open resources is probably of lower quality than an editorially compiled dictionary, which we do not have at our disposal. To check the dictionary coverage for our dataset, we used a monolingual defining dictionary for Nynorsk[2] and counted the number of meanings covered by at least one translation in our induced dictionary. Macro-averaged over all 20 lemmas, we get a coverage of 78.3%. Adjectives have the best coverage with 93%, followed by nouns with 77.7% and verbs with 57.8%.

Compared to an editorially compiled dictionary targeted at immigrants in Norway (LEXIN[3]), we see a big difference in terms of synonym selection. For the 20 lemmas, in sum our dictionary suggests 172 translations, and LEXIN suggests 84 translations, but only 43 translations are contained in both dictionaries. On the other hand, over 75% (macro-averaged over all Nynorsk lemmas) of the translations suggested by LEXIN are covered by either the same word or a synonym in our dictionary.

3.2. Corpus Analysis

The difficulty of finding the right translation in context is in part dependent on how ambiguous the word in question

[1] http://www.nrk.no/sognogfjordane/fann-ubeden-gjest-pa-sofaen-1.64529

[2] http://www.nynorskordboka.uio.no

[3] http://lexin.udir.no

Rating	Noun	Verb	Adj	All
good	21.8	14.4	15.5	17.1
acceptable	16.5	15.6	19.7	17.2
wrong	55.1	59.2	57.6	57.4
unknown translation	6.5	10.6	7.2	8.2

Table 4: Rating distribution over all 2100 rated lemma-translation pairs.

is. We have already reported the average number of translations in the previous section and will now have a look at the quality of these translations for our contexts.

The majority of translation candidates were rated as wrong in the given contexts, as shown in Table 4: On average 57.4% were deemed unsuitable for the given context, and 8.2% of the suggested translations were unknown to the annotator, leaving 34.4% that were at least acceptable. This provides another reason why post-processing of dictionary translations via CLLS is desirable.

When we look at all pairings of annotators,

$$\begin{array}{l} & good & acceptable & wrong & unknown \\ good & 162 & 175 & 166 & 40 \\ acceptable & & 98 & 315 & 34 \\ wrong & & & 809 & 219 \\ unknown & & & & 13 \end{array}$$

we see that they have chosen the exact same rating in 1082 out of 2031 cases. In 166 cases we see a contradictory rating – one annotator chose “good”, the other “wrong”. For 16 triples of word, translation and text context, the divergent rating seems to be a clear outlier compared to 3–4 consistent ratings of other annotators. These are responsible for 49 of these 166 cases of contradictory annotator pairings.

4. Method

Inspired by the results of Mihalcea (2005), we designed our system around a graph-based algorithm using undirected *PageRank* as the graph centrality algorithm. PageRank rates a node based on the ratings of connected nodes and the strength of the connection, which is approximated by the similarity between the words the two nodes represent. Starting from an uniform distribution of node weights, the rating process is iterated until scores converge. In doing so, better translations of both the target and the context words will in the optimal case be more densely connected and also have higher edge weights than translations inappropriate for the given text, leading to higher node weights for these words.

The workflow of our system is illustrated in Figure 1. To construct the graph, the text in Nynorsk first gets annotated with lemmas and PoS tags using the Oslo-Bergen tagger (Hagen et al., 2000), and each lemma is associated with all possible translation candidates from our dictionary. In the second step, all non-content words are filtered out, i.e. we consider only nouns, verbs, adjectives and adverbs both as target words and context words. Then we span a context window around the target word and add all the English translation candidates of the words inside the context window to a graph while remembering their relative position in the context window. In the graph, nodes are connected by undirected edges only when their relative position in the context window is within a certain range. As the edge weight, a similarity measure between two words is used. In Section 5, we will look at three possible implementations for this measure. When the score obtained is zero, the edge is removed. In the final graph, some nodes might become unreachable. Finally, after running the PageRank algorithm, the node scores of the translation candidates of the target word are transformed into a ranking, which is the final result.

Let us revisit our previous Nynorsk example sentence for the word *dyne* (relevant context words are underlined): *Mannen tidleg i 20-åra hadde pakka seg godt inn og lagt*

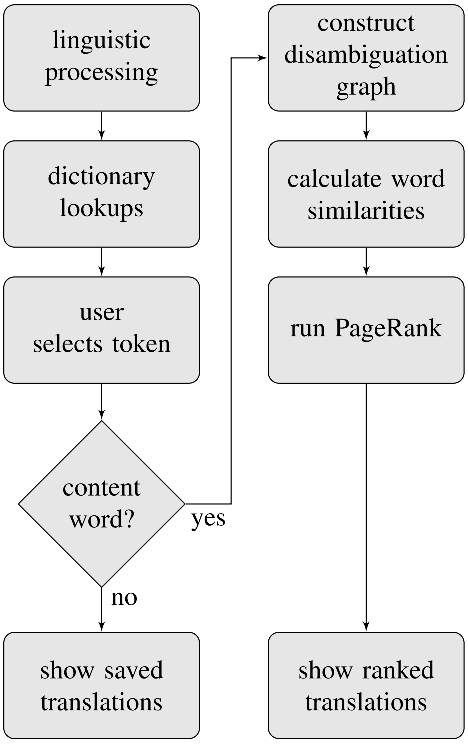

Figure 1: Workflow of the CLLS system

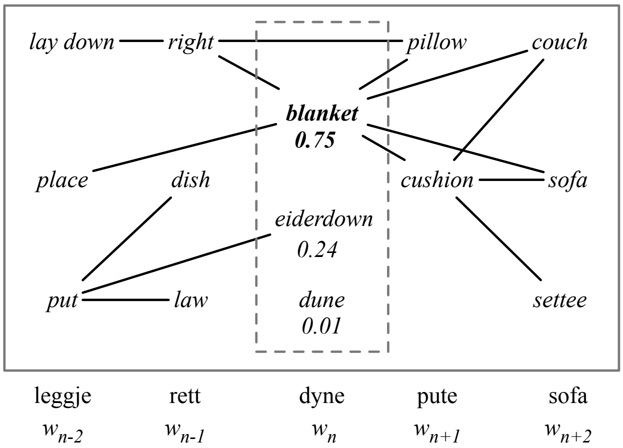

Figure 2: Disambiguation graph for the word *dyne*

	gold-scan				gold-all			
Algorithm	**Best**	**Best 3**	**Mode 3**	ρ	**Best**	**Best 3**	**Mode 3**	ρ
Upper Bound	85.4	54.9	98.0	0.97[‡]	85.5	56.4	100.0	1.00
Random Baseline	31.1	30.5	64.0	0.00	31.5	31.9	58.0	0.00
Frequency Baseline	38.3	**38.9**	**76.0**	0.24	39.2	**40.0**	**70.0**	0.24
Dictionary	**52.8**	38.2	74.0	**0.33**	**54.5**	38.9	65.0	**0.35**
PageRank Lesk	37.3	36.8	72.0	0.23	39.1	38.1	68.0	0.23
PageRank Co-occurrence	43.3	**38.4**	**75.0**	**0.24**	44.1	**39.7**	**71.0**	0.25
PageRank PMI	**45.5**	37.8	**75.0**	**0.24**	**45.8**	39.5	70.0	**0.26**

[‡] One instance of each word class had no ranking in the gold standard (either all good or all bad) and was therefore given a score of 0.0 as a fallback

Table 5: Average results on the respective gold standards. Scores for *best*, *best 3*, and *mode 3* are in percent notation.

seg til rette med ***dyne*** *og pute på sofaen til Eli.* Figure 2 exemplifies how the word *dyne* together with its context is transformed into a graph. We can see that the most appropriate translation is ranked highest because of the better contextual support.

5. Evaluation

In order to compare with the results obtained by naïvely using the same approach for computing the similarity measure for our CLLS task as Mihalcea (2005) did for WSD, we included a version of our system with a modified Lesk on glosses in the English Wiktionary as the similarity measure *PageRank Lesk*. Since Sinha and Mihalcea (2007) report improvements when incorporating other similarity measures, we use two additional approaches for our system: the similarity measure for *PageRank Co-occurrence* is based on co-occurrence statistics of PoS tagged lemmas in the Annotated English Gigaword corpus (Napoles et al., 2012); for *PageRank PMI* it is based on pointwise mutual information obtained from the same corpus.

The system has been tested using different parameter settings, both for the number of context words included in the graph and for the window size for co-occurrences. Here we only report the best of each class. For systems with corpus-based similarity measures, we have chosen a relatively big context window (up to 12 context words to the left and right of the target) for the selection of words to include in the disambiguation graph. For the edges between nodes, a small distance performs better, so we have chosen to connect each node with all the nodes within a distance of up to two positions in the context window. This value for the positional distance is also used as the window size for the construction of the models from the corpus. Lesk has different constraints on the context window size and the positional distance between connected nodes; however, using the same as with the corpus-based measures would result in too sparse a graph. Instead, we report on the setup giving the best results for this measure.

5.1. Evaluation Measures

Analogous to the Sem-Eval CLLS task (Sinha et al., 2009) we use *best* and *mode* as evaluation measures, but add *Spearman's rho* as our task differs from CLLS by the fact that we rank a fixed set of possible translations instead of proposing translations from scratch. We shall also define our measures in a more fine-grained way. Instead of calling them *best* and *mode*, we thus report *best*, *best 3* and *mode 3*. As *best* we report the average over the gold standard rating for the highest-ranking system output for all Nynorsk words. *Best 3* computes for each word the average of the top 3 system outputs, instead of just the best one. *Mode 3* evaluates the percentage of words for which the system's top three translations contained the best gold standard translation, simulating that a language learner will probably not read all possible translations but rather stop after reading the first three.

5.2. Baselines and Upper Bound

We compare our system to two baselines: a random baseline, where all translations are assigned the same confidence and therefore are ranked randomly, and a frequency baseline, where the translations are ranked according to their lemma frequency in the Annotated English Gigaword corpus (Napoles et al., 2012). We also report the results using the output order of an editorially compiled dictionary. As described in Section 3.1, only a small subset of possible translations is actually contained in both the editorially compiled dictionary and in the dictionary used for the gold standard. Because of this data mismatch, we do not consider it a proper baseline.

The upper bound reported in the tables describes the value that would be obtained if the algorithm ranked exactly like the gold standard.

5.3. Results

Table 5 shows the results for each of our gold standards. We can see that our methods are not only able to beat the random baseline, but also the frequency-based baseline in terms of *best* and *Spearman*. For these metrics, the performance of using the output order of an editorially compiled dictionary cannot be achieved by our systems, though. Comparing our different system configurations, both co-occurrence and PMI-based similarity measures clearly outperform Lesk with an advantage for PMI on the overall rank

Word Class	Best	Best 3	Mode 3	ρ
Noun	7.7	23.8	25.00	19.4
Verb	36.1	27.7	23.5	24.0
Adjective	39.6	41.0	38.5	40.9
All	26.5	31.2	28.6	26.4
Noun	5.9	2.6	-12.5	0.3
Verb	13.8	-19.6	-8.3	-1.3
Adjective	28.2	8.5	20.0	16.2
All	14.2	-3.0	0.0	3.4
Noun	-24.2	-29.5	0.0	-23.9
Verb	-35.0	11.0	18.8	-15.3
Adjective	-27.3	16.2	20.0	6.3
All	-28.2	3.4	14.3	-13.5

Table 6: Error reduction in % using PageRank PMI compared to the random baseline (top), frequency baseline (middle) and dictionary (bottom) on *gold-all.*

correlation.

Table 6 shows the result for our PMI-based scorer partitioned by word class. For the sake of better comparability, we report the relative reduction of error compared to our baselines ($\frac{score-baseline}{upper\_bound-baseline}$) instead of absolute values. Compared to the random baseline, we improve by between 26 and 30% depending on the used measure. We see that nouns are harder to improve than verbs and adjectives for the *best* metric, which we speculate is due to an on-average lower number of translations leading to a higher baseline. This is confirmed by the comparison with the frequency baseline, where nouns actually improve more than verbs. Here we can also see that verbs are mainly responsible for not beating the frequency baseline for *best 3*.

6. Comparison with Related Work

In order to be able to compare our system to the state-of-the-art described in the literature, we have also evaluated it on the dataset of the CLLS task of SemEval-2010 (Mihalcea et al., 2010). While the participants in the original task could draw on all available resources for the two well-resourced languages English and Spanish, we reach competitive results even with our low-resource approach.

For the translation from English to Spanish we use a parsed version of the English Wiktionary with JWKTL (Zesch et al., 2008). In order to evaluate with improved dictionary coverage, we evaluate a second system with a combined dictionary of Wiktionary and the data from the dictionary baseline shipped with the task evaluation data, ignoring all ranking information that came with it. We will refer to these systems as *PageRank-W* and *PageRank-X* respectively.

As the similarity measure for the edge weights, we use simple co-occurrences based on the Spanish part of the annotated Wikicorpus (Reese et al., 2010), of less than 120 million tokens in size. In contrast to the Nynorsk dataset, the Sem-Eval dataset consists of only one context sentence per test instance, which means we cannot make use of a large context window. Therefore we choose a window size of 6, which also worked comparatively well on the Nynorsk dataset. For the edges between nodes, we keep the node distance the same as with the evaluation on the Nynorsk dataset and connect nodes not more than two positions apart in the context window.

The task is different than our original evaluation in that we need to select the output translations instead of returning all translation candidates in the right order. For the *best* evaluation, we only return the top-ranked translation candidate whenever possible. When there is no unique best, we return up to 3 translation candidates with the same internal score. For the *oot* evaluation, we return the translation candidates on rank 1 to 10. If there are less than 10 candidates, we fill the rest of the slots with duplicates of the top ranked candidate.

6.1. Evaluation Measures

The SemEval CLLS task (Mihalcea et al., 2010) defines two evaluation measures: *best* and *oot* (out of ten), both accompanied by a *mode* score. The gold standard for each test instance is a multiset of translations suggested by annotators. The *mode* denotes the translation with the highest occurrence in this multiset. The score of a translation provided by the system is the share of occurrences of this translation in the multiset.

For *best*, the sum of the scores for each system translation is divided by the number of system responses, the accompanying *mode* score is the percentage of test instances where the *mode* was placed at rank 1 by the system. For *oot*, the scores of up to ten system responses are added and duplicates are allowed. The respective *mode* score is the percentage of test instances where the *mode* could be found among the ten system responses. Details can be found in the SemEval task paper.

6.2. Baselines and Upper Bound

The baselines use translations from an online English–Spanish dictionary[4]. DICT takes the first 10 translations provided by the online interface of the dictionary for the *oot* metric and the first translation for the *best* metric. For DICTCORP, all translations were retrieved and ranked according to their frequency in a corpus obtained from the Spanish Wikipedia. For the two scoring metrics, the first or up to ten translations, respectively, are considered (Mihalcea et al., 2010).

The upper bound for *best* is reported to be 40.57, for *oot* 405.78 (Mihalcea et al., 2010).

6.3. Results

Tables 7 and 8 show the results for the evaluation on the SemEval dataset. Our original system (PageRank-W) ends up slightly below the ranked dictionary baseline and about 10 percent points behind the top scorer for the *best mode* metric. For the *oot mode* metric we see the same, yet here we are ahead of all other systems when considering the *oot* score. However, the evaluation output shows duplicates for 920 out of 1000 test instances, which means *oot mode* mostly judged our dictionary, not our system.

With the extended dictionary (PageRank-X), on the other hand, we are outperformed by only one competitor of the

[4] `http://spanishdict.com`

System	R	P	Mode R	Mode P
1. USPwlv	26.81	26.81	58.85	58.85
2. UBA-T	27.15	27.15	57.29	57.20
3. ColSlm	25.99	27.99	56.24	59.16
4. WLVusp	25.27	25.27	52.81	52.81
5. † DICT	24.34	24.34	50.34	50.34
6. **PageRank-W**	23.14	23.61	48.15	49.30
7. SWAT-E	21.46	21.46	43.21	43.21
...				
14. **PageRank-X**	17.40	17.40	34.71	34.71
15. IRST-1	15.38	22.16	33.47	40.04
16. † DICTCORP	15.09	15.09	29.22	29.22
...				

Table 7: **best** recall (R) and precision (P) for a selection of systems ranked after *mode R*. The baselines are marked with †, our own systems are in ***bold***.

System	oot R	oot P	Mode R	Mode P
1. UBA-W	52.75	52.75	83.54	83.54
2. **PageRank-X**	102.61	102.61	82.44	82.44
3. UBA-T	47.99	47.99	81.07	81.07
4. USPwlv	47.60	47.60	79.84	79.84
...				
7. UvT-g	55.29	55.19	73.94	73.94
8. † DICT	44.04	44.04	73.53	73.53
9. **PageRank-W**	204.33	208.50	72.84	74.58
10. † DICTCORP	42.65	42.65	71.60	71.60
11. ColEur	41.72	44.77	67.35	71.47
12. SWAT-E	174.59	174.59	66.94	66.94
...				

Table 8: **oot** recall (R) and precision (P) for a selection of systems ranked after *mode R*. The baselines are marked with †, our own systems are in ***bold***.

original task for *oot* and its *mode*, respectively. For *oot mode* we are only about one percentage point behind. However, extending the dictionary is detrimental for the *best* evaluation as there are more possible translation candidates to rank, but we still outperform the frequency baseline.

In order to determine the respective influence of dictionary and ranking system, we compare our system to a random order of the dictionary translations, using two different dictionaries (see Table 9). After the ranking is determined, the selection of the output is the same for the random order as for our system. One dictionary is the Wiktionary-based one used for PageRank-W. The other one we constructed from the gold standard for this task. For each lemma we collected the union of all annotator responses irrespective of the context and used them as the unranked dictionary output. This way we can be sure the *mode* is actually among the translation candidates. The gold standard translations are normalized, i.e. they lack all diacritics, so we automatically inflate them to make sure the lemma in its correct form is among the translation candidates. This adds some additional noise for our system, but will not influence the random order. Note that we still need to use our Wiktionary-based dictionary for all context words not covered by the dictionary-based on the gold standard.

System	Best	Mode	oot	Mode	Mode 3
PageRank-W	23.14	48.15	204.33	72.84	69.41
Random-W	16.62	34.43	168.55	72.29	60.61
PageRank-G	10.13	14.95	70.88	83.26	41.98
Random-G	5.55	6.45	54.58	58.57	21.81

Table 9: **best** (left), **oot** (middle), and **mode 3** (right) recall for different dictionaries (W for Wiktionary, G for gold-dictionary) for our system and a random order of dictionary entries.

Our system improves significantly over the random order of either dictionary. The exception is the *oot mode* score for the Wiktionary dictionary, which is limited by the number of provided translation candidates. Overall, using a bigger dictionary to increase coverage works better for the *oot* metric, while a more restricted dictionary focusing on prominent translations works better for the *best* metric.

7. Discussion and Conclusion

The evaluations on our own dataset and on the CLLS task show a similar picture. An editorially compiled ranked dictionary is hard to beat, but when only dictionaries without reliable ranking information are available, a graph-based approach makes it possible to obtain similar results both for high-resource and low-resource languages. When our system can rely on a dictionary of reasonable quality, only few other systems get better results even for high-resource languages. We have shown that for low-resource languages, cross-lingual lexical substitution can also give an advantage over a context-free ranking using lemma frequency, despite the lack of resources typically used by other CLLS systems.

While the evaluation results from Section 5 are not directly comparable to other results found in the literature, we will try to set them at least in perspective. In a comparison of algorithms for graded word sense disambiguation, Friedrich et al. (2012) report a ρ of 0.210 on the WSsim-1 test set for a reimplementation of the PageRank-based system presented by Sinha and Mihalcea (2007) with extended Lesk as the similarity measure. Our results are slightly higher, but our Lesk-based variant is quite close to the aforementioned.

In order to be able to set our results in relation to the results obtained for the Sem-Eval 2010 CLLS task on the language pair English–Spanish, we calculate the error reduction for listed systems using their reported baselines and obtainable best for the *best* measure. We obtained an improvement of 26.5% and 14.2% compared to our random and frequency baseline, respectively. Compared to a frequency baseline, all but one system attending the CLLS task could improve between 14.8% and 49.1%. Compared to a ranked dictionary baseline, however, only four systems could improve between 5.7% and 20.0%.

In a post-hoc analysis, we constructed our own ranked dictionary baseline, but cannot even come close to it in the *best* measure. On the other hand, we see an improvement of 14.3% compared to this dictionary for *mode 3*, which, however, we cannot directly compare to the Sem-Eval systems. Testing our system directly on the Sem-Eval dataset, we see a similar picture. We are clearly ahead of the frequency baseline, but cannot outperform the ranked dictionary baseline, although we are close. When we increase the number of translation candidates provided by our dictionary, our system falls further behind, much like on the Nynorsk dataset, where we also combined several lexical resources. Yet the filtering capabilities of our system for finding the top 10 translation candidates work especially well with an expanded dictionary.

A comparison with results obtained with an unrealistic gold-dictionary, i.e. a dictionary artificially constructed for 100% coverage on the gold standard, reveals that the selection of translation candidates indeed has a substantial effect on the result. For all metrics but *oot mode*, even a random ranking of the more restrictive dictionary gives better results than our system output on this gold-dictionary. We must assume that the gold-dictionary behaves more like a dictionary induced from a parallel corpus than a typical context-free dictionary. It is likely to include translation pairs that would normally not be considered translations of each other, for example when a word is used as part of a multiword expression in one of the context sentences. Such translation candidates would be distractors for other contexts, as they could very well be the best translation in context for a different word in the source language. At the scoring step, the information about the word used in the source language cannot be used in our approach, as we do not have any linking information besides the dictionary, hence the need for a rather selective approach to candidate collection.

To conclude, we built a gold standard with 100 annotated instances for the language pair Nynorsk–English. Our results on this gold standard indicate that the chosen approach is feasible even for low-resource languages. For verbs, however, the system does not perform as well as a naïve frequency baseline, except on the *best* metric. In future work we will further inspect the data to find out why this is the case. We have also shown that this approach can be generalized to other language pairs and that we get competitive results comparable to systems with higher resource demands. We get close to a ranked dictionary baseline, and the gap to the best systems is not very large. The selection of the dictionary has a big influence on the result, however, which suggests that the candidate collection process needs to get a stronger focus in future work. Our results reveal that the open lexical resources we used work fairly well in that regard.

8. Acknowledgements

The authors wish to thank the anonymous annotators of the dataset. This work was funded by the Cluster of Excellence “Multimodal Computing and Interaction” of the German Excellence Initiative.

9. Bibliographical References

Ananiadou, S., McNaught, J., Thompson, P., Rehm, G., and Uszkoreit, H. (2012). *The English Language in the Digital Age*. Springer.

Apidianaki, M. (2011). Unsupervised cross-lingual lexical substitution. In *Proceedings of the first workshop on unsupervised learning in NLP*, pages 13–23. Association for Computational Linguistics.

Basile, P. and Semeraro, G. (2010). Uba: Using automatic translation and wikipedia for cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 242–247, Uppsala, Sweden, July. Association for Computational Linguistics.

Friedrich, A., Engonopoulos, N., Thater, S., and Pinkal, M. (2012). A comparison of knowledge-based algorithms for graded word sense assignment. In Martin Kay et al., editors, *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Posters, 8-15 December 2012, Mumbai, India*, pages 329–338. Indian Institute of Technology Bombay.

Hagen, K., Johannessen, J. B., and Nøklestad, A. (2000). A Constraint-Based Tagger for Norwegian. In Carl-Erik Lindberg et al., editors, *17th Scandinavian Conference of Linguistics*, volume 19 of *Odense Working Papers in Language and Communication*, pages 31–48. Syddansk Universitet, Odense.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Maurud, Ø. (1976). Reciprocal comprehension of neighbour languages in scandinavia: An investigation of how well people in denmark, norway and sweden understand each other’s written and spoken languages. *Scandinavian journal of educational research*, 20(1):49–72.

McCarthy, D. and Navigli, R. (2009). The english lexical substitution task. *Language resources and evaluation*, 43(2):139–159.

McCarthy, D., Sinha, R., and Mihalcea, R. (2013). The cross-lingual lexical substitution task. *Language resources and evaluation*, 47(3):607–638.

Mihalcea, R., Sinha, R., and McCarthy, D. (2010). Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval ’10, pages 9–14, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mihalcea, R. (2005). Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*, pages 411–418. The Association for Computational Linguistics.

Napoles, C., Gormley, M., and Van Durme, B. (2012). Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100. Association for Computational Linguistics.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Reese, S., Boleda, G., Cuadros, M., Padró, L., and Rigau, G. (2010). Wikicorpus: A word-sense disambiguated multilingual wikipedia corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Sinha, R. S. and Mihalcea, R. (2007). Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007), September 17-19, 2007, Irvine, California, USA*, pages 363–369. IEEE Computer Society.

Sinha, R., McCarthy, D., and Mihalcea, R. (2009). Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, DEW '09, pages 76–81, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sinha, R. S. (2013). *Finding Meaning in Context Using Graph Algorithms in Mono- and Cross-lingual Settings*. Ph.D. thesis, University of North Texas, Denton, TX, USA. AAI3579244.

Unhammer, K. B. and Trosterud, T. (2009). Reuse of Free Resources in Machine Translation between Nynorsk and Bokmål. In Juan Antonio Pérez-Ortiz, et al., editors, *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 35–42, Alicante. Universidad de Alicante. Departamento de Lenguajes y Sistemas Informáticos.

Vilarino, D., Balderas, C., Pinto, D., Rodríguez, M., and León, S. (2010). Fcc: Modeling probabilities with giza++ for task# 2 and# 3 of semeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 112–116. Association for Computational Linguistics.

Wicentowski, R., Kelly, M., and Lee, R. (2010). Swat: Cross-lingual lexical substitution using local context matching, bilingual dictionaries and machine translation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 123–128, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zahran, M. A., Raafat, H., and Rashwan, M. (2015). Cross lingual lexical substitution using word representation in vector space. In *The Twenty-Eighth International Flairs Conference*.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

10. Language Resource References

Norsk Ordbok 2014. (2012). *Norsk Ordboks Nynorskkorpus*. University of Oslo via http://no2014.uio.no/korpuset/.