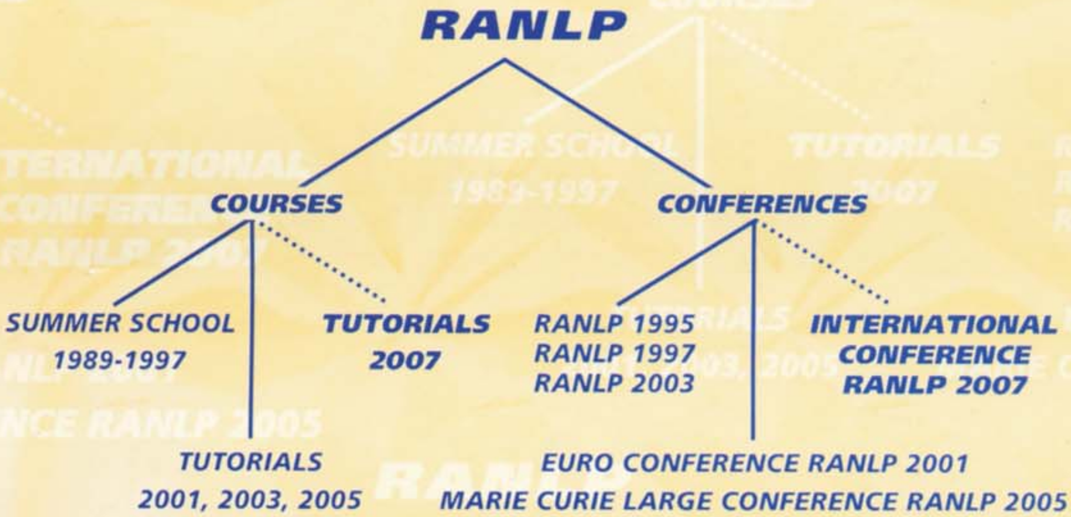
INTERNATIONAL CONFERENCE
RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING

PROCEEDINGS

Edited by
Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, Nikolai Nikolov


RANLP

COURSES

CONFERENCES

SUMMER SCHOOL 1989-1997

TUTORIALS 2007

TUTORIALS 2001, 2003, 2005

RANLP 1995
RANLP 1997
RANLP 2003

INTERNATIONAL CONFERENCE RANLP 2007

EURO CONFERENCE RANLP 2001
MARIE CURIE LARGE CONFERENCE RANLP 2005

27-29 September 2007, Borovets, Bulgaria

XEROX
Research Centre Europe

BIS 21++

[Nakov & колектив, 2007a] Nakov P., Nakov S., Paskaleva E. "Improved Word Alignments Using the Web as a Corpus", *Proceedings of International Conference "Recent Advances in Natural Language Processing" (RANLP 2007)*, pages 400-405, Borovets, Bulgaria, 2007

Improved Word Alignments Using the Web as a Corpus

Preslav Nakov
UC Berkeley
EECS, CS division
Berkeley, CA 94720
nakov@cs.berkeley.edu

Svetlin Nakov
Sofia University
5 James Boucher Blvd.
Sofia, Bulgaria
nakov@fmi.uni-sofia.bg

Elena Paskaleva
Bulgarian Academy of Sciences
25A Acad. G. Bonchev Str.
Sofia, Bulgaria
hellen@lml.bas.bg

Abstract

We propose a novel method for improving word alignments in a parallel sentence-aligned bilingual corpus based on the idea that if two words are translations of each other then so should be many words in their local contexts. The idea is formalised using the Web as a corpus, a glossary of known word translations (dynamically augmented from the Web using bootstrapping), the vector space model, linguistically motivated weighted minimum edit distance, competitive linking, and the IBM models. Evaluation results on a Bulgarian-Russian corpus show a sizable improvement both in word alignment and in translation quality.

Keywords

Machine translation, word alignments, competitive linking, Web as a corpus, string similarity, edit distance.

1 Introduction

The beginning of modern *Statistical Machine Translation* (*SMT*) can be traced back to 1988, when Brown et al. [5] from IBM published a formalised mathematical formulation of the translation problem and proposed five word alignment models – IBM models 1, 2, 3, 4 and 5. Starting with a bilingual parallel sentence-aligned corpus, the IBM models learn how to translate individual words and the probabilities of these translations. Later, decoders like the ISI ReWrite Decoder [9] became available, which made it possible to quickly build SMT systems with decent quality.

An important shift happened in 2004, when the Pharaoh model [11] has been proposed, which uses whole *phrases* (typically of length up to 7, not necessarily representing linguistic units), rather than just words. This led to a significant improvement in translation quality, since phrases can encode local gender/number agreement, facilitate choosing the correct sense for ambiguous words, and naturally handle fixed phrases and idioms. While methods have been proposed for learning translation phrases directly [17], the most popular *alignment template approach* [23] requires bi-directional word alignments at the sentence level from which phrases consistent with those alignments are extracted. Since better word alignments can lead to better phrases[1], improving word alignments remains one of the primary research problems in SMT: in fact, there are more papers published yearly on word alignments than on any other SMT subproblem.

In the present paper, we describe a novel method for improving word alignments using the Web as a corpus, a glossary of known word translations (dynamically augmented from the Web using bootstrapping), the vector space model, weighted minimum edit distance, competitive linking, and the IBM models. The potential of the method is demonstrated on a Bulgarian-Russian bilingual corpus.

The rest of the paper is organised as follows: section 2 explains the method in detail, section 3 describes the corpus and the resources used, section 4 contains the evaluation, section 5 points to important related research, and section 6 concludes with some possible directions for future work.

2 Method

Our method combines two similarity measures which make use of different information sources. First, we define a language-specific modified minimum edit distance, based on linguistically-motivated rules targeting Bulgarian-Russian cognate pairs. Second, we define a distributional semantic similarity measure, based on the idea that if two words represent a translations pair, then the most frequently co-occurring words in their local contexts should be translations of each other as well. This intuition is formalised using the Web as a corpus, a bilingual glossary of word translation pairs used as "bridges", and the vector space model. The two measures are combined with competitive linking [19] in order to obtain high quality word translation pairs, which are then appended to the bilingual sentence-aligned corpus in order to bias the subsequent training of the IBM word alignment models [5].

2.1 Orthographic Similarity

We use an orthographic similarity measure, which is based on the *minimum edit distance* (MED) or Levenshtein distance [16]. MED calculates the distance between two strings s_1 and s_2 as the minimum number of edit operations – INSERT, REPLACE, DELETE – needed to transform s_1 into s_2. For example, the MED between *r.* **первый** (Russian, '*first*') and *b.* **първият**

[1] The dependency between word alignments and translation quality is indirect; improving the former does not necessarily improve the latter.

(Bulgarian, '*the first*') is 4: three REPLACE operations (**е** → **ъ**, **ы** → **и**, **й** → **я**) and one INSERT (of **т**).

We modify the classic MED in two ways. First, we normalise the two strings, taking into account some general graphemic correlations between the phonetico-graphemic systems of the two closely-related Slavonic languages – Bulgarian and Russian:

- For Russian words, we remove the letters **ь** and **ъ**, as their graphemic collocations are excluded in Bulgarian, e.g. **ь** between two consonants (*r.* **сильно** ↔ *b.* **силно**, *strongly*), **ъ** following a consonant (*r.* **объявление** ↔ *b.* **обявление**, *an announcement*), etc.

- For Russian words, we remove the ending **й**, which is the typical nominative adjective ending in Russian, but not in Bulgarian, e.g. *r.* **детский** ↔ *b.* **детски** (*children's*).

- For Bulgarian words, we remove the definite article, e.g. *b.* **горският** (*the forestal*) → *b.* **горски** (*forestal*). The definite article is the only agglutinative morpheme in Bulgarian and has no counterpart in Russian: Bulgarian has definite, but not indefinite article, and there are no articles in Russian.

- We transliterate the Russian-specific letters (missing in the Bulgarian alphabet) or letter combinations in a regular way: **ы** ↔ **и**, **э** ↔ **е**, and **шт** ↔ **щ**, e.g. *r.* **электрон** ↔ *b.* **електрон** (*an electron*), *r.* **выл** ↔ *b.* **вил** (past participle of *to howl*), *r.* **штаб** ↔ *b.* **щаб** (mil. *a staff*), etc.

- Finally, we remove all double letters in both languages (e.g. **нн** → **н**; **сс** → **с**): While consonant and vowel doubling is very rare in Bulgarian (except at morpheme boundaries for a limited number of morphemes), it is more common in Russian, e.g. in case of words of foreign origin: *r.* **ассамблея** → *b.* **асамблея** (*an assembly*)

Second, we use different letter-pair specific costs for REPLACE. We use 0.5 for all vowel to vowel substitutions, e.g. **о** ↔ **е** as in *r.* **лицо** ↔ *b.* **лице** (*a face*). We also use 0.5 for some consonant-consonant replacements, e.g. **с** ↔ **з**. Such regular phonetic changes are reflected in different ways in the orthographic systems of the two languages, Bulgarian being more conservative and sticking to morphological principles. For example, in Bulgarian the final **з** in prefixes like **из-** and **раз-** never change to **с**, while in Russian they sometimes do, e.g. *r.* **исследователь** ↔ *b.* **изследовател** (*an explorer*), *r.* **рассказ** ↔ *b.* **разказ** (*a story*), etc.

We use a cost of 1 for all other replacements.

It is easy to see that this *modified minimum edit distance* (MMED) is more adequate than MED – it is only 0.5 for *r.* **первый** and *b.* **първият**: we first normalise them to **перви** and **първи**, and then we do a single vowel-vowel REPLACE with the cost of 0.5.

We transform MMED into a similarity measure, *modified minimum edit distance ratio* (MMEDR) using the following formula ($|s|$ is the number of letters in s before the normalisation):

$$\text{MMEDR}(s_1, s_2) = 1 - \frac{\text{MMED}(s_1,s_2)}{\max(|s_1|,|s_2|)}$$

Below we compare MMEDR with *minimum edit distance ratio* (MEDR):

$$\text{MEDR}(s_1, s_2) = 1 - \frac{\text{MED}(s_1,s_2)}{\max(|s_1|,|s_2|)}$$

and *longest common subsequence ratio* (LCSR) [18]:

$$\text{LCSR}(s_1, s_2) = \frac{|\text{LCS}(s_1,s_2)|}{\max(|s_1|,|s_2|)}$$

In the last definition, LCS(s_1, s_2) refers to the longest common subsequence of s_1 and s_2, e.g. LCS(**первый**, **първият**) = **прв**, and therefore

MMEDR(**первый**, **първият**) = 3/7 ≈ 0.43

We obtain the same score using MMED:

MMED(**первый**, **първият**) = 1 − 4/7 ≈ 0.43

while with MMEDR we have:

MMEDR(**первый**, **първият**) = 1 − 0.5/7 ≈ 0.93

2.2 Semantic Similarity

The second basic similarity measure we use is WEB-ONLY, which measures the semantic similarity between a Russian word w_{ru} and a Bulgarian word w_{bg} using the Web as a corpus and a glossary G of known Bulgarian-Russian translation pairs used as "bridges". The basic idea is that if two words are translations of each other then many of the words in their respective local contexts should be mutual translations as well.

First, we issue a query to Google for w_{ru} or w_{bg}, limiting the language to Russian or Bulgarian, and we collect the text from the resulting 1,000 snippets. We then extract the words from the local context (two words on either side of the target word), we remove the stopwords (prepositions, pronouns, conjunctions, interjections and some adverbs), we lemmatise the remaining words, and we filter out the words that are not in G. We further replace each Russian word with its Bulgarian counter-part in G. As a result, we end up with two Bulgarian frequency vectors, corresponding to w_{ru} and w_{bg}, respectively. Finally, we TF.IDF-weight the vector coordinates [31] and we calculate the semantic similarity between w_{bg} and w_{ru} as the cosine between their corresponding vectors.

2.3 Combined Similarity Measures

In our experiments (see below), we have found that MMEDR yields a better precision, while WEB-ONLY has a better recall. Therefore we tried to combine the two similarity measures in different ways:

- WEB-AVG: *average* of WEB-ONLY and MMEDR;

- WEB-MAX: *maximum* of WEB-ONLY and MMEDR;

- WEB-CUT: The value of WEB-CUT(s_1, s_2) is 1, if MMEDR$(s_1, s_2) \geq \alpha$ $(0 < \alpha < 1)$, and is equal to WEB-ONLY(s_1, s_2), otherwise.

2.4 Competitive Linking

The above similarity measures are used in combination with *competitive linking* [19], which assumes that a source word is either translated with a single target word or is not translated at all. Given a sentence pair, the similarity between all Bulgarian-Russian word pairs is calculated[2], which induces a fully-connected weighted bipartite graph. Then a greedy approximation to the maximum weighted bipartite matching in that graph is extracted as follows: First, the most similar pair of unaligned words is aligned and both words are discarded from further consideration. Then the next most similar pair of unaligned words is aligned and the two words are discarded, and so forth. The process is repeated until there are no unaligned words left or until the maximal word pair similarity falls below a pre-specified threshold θ ($0 \leq \theta \leq 1$), which could leave some words unaligned.

3 Resources

3.1 Parallel Corpus

We use a parallel sentence-aligned Bulgarian-Russian corpus: the Russian novel *Lord of the World*[3] by Alexander Beliaev and its Bulgarian translation[4]. The text has been sentence aligned automatically using the alignment tool *MARK ALISTeR* [26], which is based on the Gale-Church algorithm [8]. As a result, we obtained 5,827 parallel sentences, which we divided into *training* (4,827 sentences), *tuning* (500 sentences), and *testing set* (500 sentences).

3.2 Grammatical Resources

We use monolingual dictionaries for lemmatisation. For Bulgarian, we use a large morphological dictionary, containing about 1,000,000 wordforms and 70,000 lemmata [25], created at the Linguistic Modeling Department, Institute for Parallel Processing, Bulgarian Academy of Sciences. The dictionary is in DELAF format [30]: each entry consists of a wordform, a corresponding lemma, followed by morphological and grammatical information. There can be multiple entries for the same wordform, in case of multiple homographs. We also use a large grammatical dictionary of Russian in the same format, consisting of 1,500,000 wordforms and 100,000 lemmata, based on the Grammatical Dictionary of A. Zaliznjak [33]. Its electronic version was supplied by the Computerised fund of Russian language, Institute of Russian language, Russian Academy of Sciences.

3.3 Bilingual Glossary

We built a bilingual glossary from an online Bulgarian-Russian dictionary[5]. First, we removed all multi-word expressions. Then we combined each Russian word with each Bulgarian one – due to polysemy/homonymy some words had multiple translations. As a result, we obtained a glossary G of 3,794 word translation pairs.

Due to the modest glossary size, in our initial experiments, we were lacking translations for many of the most frequent context words. For example, when comparing *r.* **платье** (*a dress*) and *b.* **рокля** (*a dress*), we find adjectives like *r.* **свадебное** (*wedding*) and *r.* **вечернее** (*evening*) among the most frequent Russian context words, and *b.* **сватбена** and *b.* **вечерна** among the most frequent Bulgarian context words. While missing in our bilingual glossary, it is easy to see that they are orthographically similar and thus likely cognates. Therefore, we automatically extended G with possible cognate pairs. For the purpose, we collected the most frequent 30 non-stopwords RU_{30} and BG_{30} from the local contexts of w_{ru} and w_{bg}, respectively, that were missing in our glossary. We then calculated the MMEDR for every word pair $(r, b) \in (RU_{30}, BG_{30})$, and we added to G all pairs for which the value was above 0.90. As a result, we managed to extend G with 6,289 additional high-quality translation pairs.

4 Evaluation

We evaluate the similarity measures in four different ways: manual analysis of WEB-CUT, alignment quality of competitive linking, alignment quality of the IBM models for a corpus augmented with word translations from competitive linking, and translation quality of a phrase-based SMT trained on that corpus.

4.1 Manual Evaluation of WEB-CUT

Recall that by definition WEB-CUT(s_1, s_2) is 1, if MMEDR$(s_1, s_2) \geq \alpha$, and is equal to WEB-ONLY(s_1, s_2), otherwise. To find the best value for α, we tried all values $\alpha \in \{0.01, 0.02, 0.03, \ldots, 0.99\}$. For each value, we word-aligned the training sentences from the parallel corpus using competitive linking and WEB-CUT, and we extracted a list of the distinct aligned word pairs, which we added twice as additional "sentence" pairs to the training corpus. We then calculated the perplexity of IBM model 4 for that augmented corpus. This procedure was repeated for all candidate values of α, and finally $\alpha = 0.62$ was selected as it yielded the lowest perplexity.[6]

The last author, a native speaker of Bulgarian who is fluent in Russian, manually examined and annotated as *correct*, *rough* or *wrong* the 14,246 distinct aligned Bulgarian-Russian word type pairs, obtained with competitive linking and WEB-CUT for $\alpha = 0.62$. The following groups naturally emerge:

1. **"Identical" word pairs** (MMEDR$(s_1, s_2) = 1$): 1,309 or 9% of all pairs. 70% of them are completely identical, e.g. **скоро** (*soon*) is spelled the same way in both Bulgarian and Russian. The remaining 30% exhibit regular graphemic changes, which are recognised by MMEDR (See section 2.1.)

[2] Due to their special distribution, stopwords and short words (one or two letters) are not used in competitive linking.

[3] http://www.lib.ru

[4] http://borislav.free.fr/mylib

[5] http://www.bgru.net/intr/dictionary/

[6] This value is close to 0.58, which has been found to perform best for LCSR on Western-European languages [15].

2. **“True friends”** ($\alpha \leq$ MMEDR$(s_1, s_2) < 1$): 5,289 or 37% of all pairs. This group reflects changes combining regular phonemic and morphemic (grammatical) correlations. Examples include similar but not identical affixes (e.g. the Russian prefixes **во-** and **со-** become **въ-** and **съ-** in Bulgarian), similar graphemic shapes of morpheme values (e.g. the Russian singular feminine adjective endings **-ая** and **-яя** become **-а** and **-я** in Bulgarian), etc.

3. **“Translations”** (MMEDR$(s_1, s_2) < \alpha$): 7,648 or 54 % of all pairs. Here the value of WEB-ONLY(s_1, s_2) is used. We divide this group into the following sub-categories: *correct* (73%), *rough* (3%) and *wrong* (24%).

Our analysis of the *rough* and *wrong* sub-groups of the latter group exposes the inadequacy of the idea of reducing sentence translation to a sequence of word-for-word translations, even for closely related languages like Bulgarian and Russian. Laying aside the translator’s freedom of choice, the translation correspondences often link a word to a phrase, or a phrase to another phrase, often idiomatically, and sometimes involve syntactic transformations as well. For example, when aligning the Russian word *r.* **отвернуться** to its Bulgarian translation *b.* **обръщам гръб** (*to turn back*), competitive linking wrongly aligns *r.* **отвернуться** to *b.* **гръб** (*a back*). Similarly, when the Russian for *to challenge*, *r.* **бросать вызов** (lit. *to throw a challenge*), is aligned to its Bulgarian translation *b.* **хвърлям ръкавица** (lit. *to throw a glove*), this results in wrongly aligning *r.* **вызов** (*a challenge*) to *b.* **ръкавица** (*a glove*). Note however that such alignments are still helpful in the context of SMT.

Figure 1 shows the precision-recall curve for the manual evaluation of competitive linking with WEB-CUT for the third group only (MMEDR$(s_1, s_2) < \alpha$), considering both *rough* and *wrong* as incorrect. We can see that the precision is 0.73 even for recall of 1.

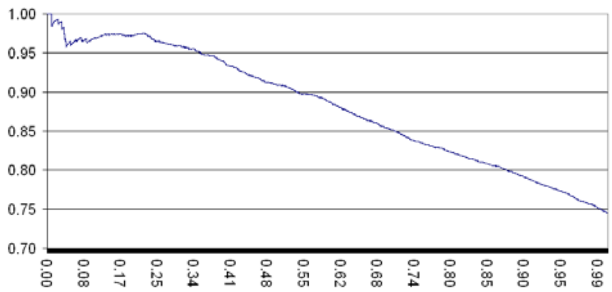


Fig. 1: Manual evaluation of WEB-CUT: *Precision-recall curve for competitive linking with* WEB-CUT *on the “translations” sub-group (*MMEDR$(s_1, s_2) < 0.62$*).*

4.2 Word Alignments

4.2.1 Gold Standard Word Alignments

The last author, a linguist, manually aligned the first 100 sentences from the training corpus, thus creating a gold standard for calculating the *alignment error rate* (*AER*) for the different similarity measures.

Manual alignments typically use two kinds of links: *sure* and *possible*. As we have seen above, even for closely related languages like Russian and Bulgarian, the alignment of each source word to a target one could be impossible, unless a suitable convention is adopted. Particularly problematic are the “hanging” single words – typically stemming from syntactic differences. We prefer to align such word to the same target word to which is aligned the word it is dependent on, and to mark the link as *possible*, rather than *sure*. More formally, if the source Russian word x_{ru} is translated with a pair of target Bulgarian words x_{bg} and y_{bg}, where x_{ru} is a *sure* translation of x_{bg}, and y_{bg} is a grammatical or “empty” word ensuring the correct surface presentation of the grammatical/lexical relation, then we add a *possible* link between y_{bg} to x_{ru} as well.

For instance, the Russian genitive case is typically translated in Bulgarian with a prepositional phrase, **на**+*noun*, e.g. *r.* **звуки музыки** (*sounds of music*) is translated as *b.* **звуците на музиката**. Other examples include regular ellipsis/dropping of elements specific for one of the languages only, e.g. subject dropping in Bulgarian, ellipsis of Russian auxiliaries in present tense, etc. For example, *r.* **я знал** (*I knew*) can be translated as *b.* **аз знаех**, but also as *b.* **знаех**. On the other hand, *r.* **он герой** (‘*he is a hero*’, lit. ‘*he hero*’) is translated as *b.* **той е герой** (lit. ‘*he is hero*’).

4.2.2 Competitive Linking

Figure 2 shows the AER for competitive linking with all 7 similarity measures: our orthographic and semantic measures (MMEDR and WEB-ONLY), the three combinations (WEB-CUT, WEB-MAX and WEB-AVG), as well as for LCSR and MEDR. We can see an improvement of up to 6 AER points when going from LCSR/MEDR/WEB-ONLY to WEB-CUT/WEB-AVG. Note that here we calculated the AER on a modified version of the 100 gold standard sentences – the stopwords and the punctuation were removed in order to ensure a fair comparison with competitive linking, which ignores them. In addition, each of the measures has its own threshold θ for competitive linking (see section 2.4), which we set by optimising perplexity on the training set, as we did for α in the section 4.1: we tried all values of $\theta \in \{0.05, 0.10, \ldots, 1.00\}$, and we selected the one which yielded the lowest perplexity.

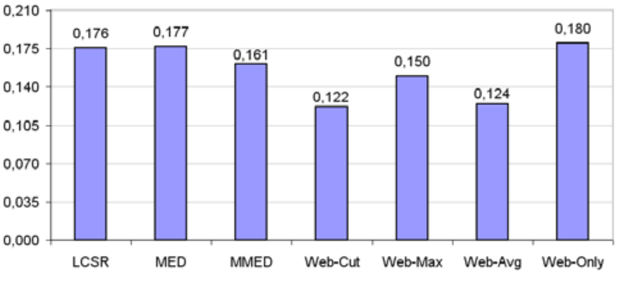


Fig. 2: AER for competitive linking: *stopwords and punctuation are not considered.*

4.2.3 IBM Models

In our next experiment, we first extracted a list of the distinct word pairs aligned with competitive linking, and we added them twice as additional "sentence" pairs to the training corpus, as in section 4.1. We then generated two directed IBM model 4 word alignments (Bulgarian → Russian, Russian → Bulgarian) for the new corpus, and we combined them using the *interect+grow heuristic* [22]. Table 3 shows the AER for these combined alignments. We can see that while training on the augmented corpus lowers AER by about 4 points compared to the baseline (which is trained on the original corpus), there is little difference between the similarity measures.

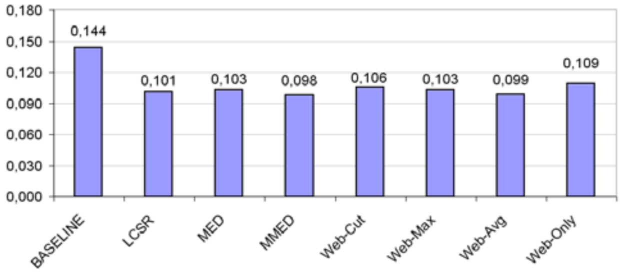


Fig. 3: AER for IBM model 4: *intersect+grow.*

4.3 Machine Translation

As we said in the introduction, word alignments are an important first step in the process of building a phrase-based SMT. However, as many researchers have reported, better AER does not necessarily mean improved machine translation quality [2]. Therefore, we built a full Russian → Bulgarian SMT system in order to assess the actual impact of the corpus augmentation (as described in the previous section) on the translation quality.

Starting with the symmetrised word alignments described in the previous section, we extracted phrase-level translation pairs using the *alignment template approach* [13]. We then trained a log-linear model with the standard feature functions: language model probability, word penalty, distortion cost, forward phrase translation probability, backward phrase translation probability, forward lexical weight, backward lexical weight, and phrase penalty. The feature weights, were set by maximising Bleu [24] on the development set using *minimum error rate training* [21].

Tables 4 and 5 show the evaluation on the test set in terms of Bleu and NIST scores. We can see a sizable difference between the different similarity measures: the combined measures (WEB-CUT, WEB-MAX and WEB-AVG) clearly outperforming LCSR and MEDR. MMEDR outperforms them as well, but the difference from LCSR is negligible.

5 Related Work

Many researchers have exploited the intuition that words in two different languages with similar or identical spelling are likely to be translations of each other.

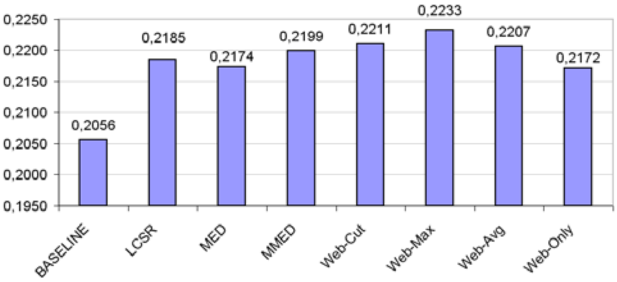


Fig. 4: Translation quality: *Bleu score.*

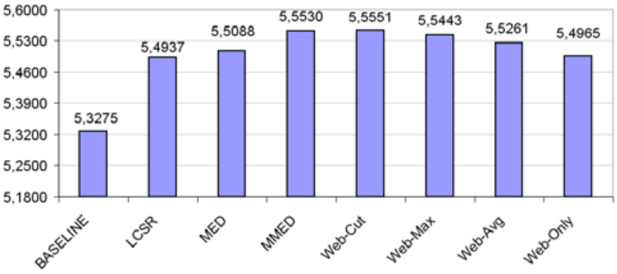


Fig. 5: Translation quality: *NIST score.*

Al-Onaizan & al. [1] create improved Czech-English word alignments using probable cognates extracted with one of the variations of LCSR [18] described in [32]. They tried to constrain the co-occurrences, to seed the parameters of IBM model 1, but their best results were achieved by simply adding the cognates to the training corpus as additional "sentences". Using a variation of that technique, Kondrak, Marcu and Knight [15] demonstrated improved translation quality for nine European languages. We extend this work, by adding competitive linking [19], language-specific weights, and a Web-based semantic similarity measure.

Koehn & Knight [12] describe several techniques for inducing translation lexicons. Starting with unrelated German and English corpora, they look for (1) identical words, (2) cognates, (3) words with similar frequencies, (4) words with similar meanings, and (5) words with similar contexts. This is a bootstrapping process, where new translation pairs are added to the lexicon at each iteration.

Rapp [27] describes a correlation between the co-occurrences of words that are translations of each other. In particular, he shows that if in a text in one language two words A and B co-occur more often than expected by chance, then in a text in another language the translations of A and B are also likely to co-occur frequently. Based on this observation, he proposes a model for finding the most accurate cross-linguistic mapping between German and English words using non-parallel corpora. His approach differs from ours in the similarity measure, the text source, and the addressed problem. In later work on the same problem, Rapp [28] represents the context of the target word with four vectors: one for the words immediately preceding the target, another one for the ones immediately following the target, and two more for the words one more word before/after the target.

Fung and Yee [7] extract word-level translations from non-parallel corpora. They count the number of sentence-level co-occurrences of the target word with a fixed set of “seed” words in order to rank the candidates in a vector-space model using different similarity measures, after normalisation and TF.IDF-weighting [31]. The process starts with a small initial set of seed words, which are dynamically augmented as new translation pairs are identified. We do not have a fixed set of seed words, but generate it dynamically, since finding the number of co-occurrences of the target word with each of the seed words would require prohibitively many search engine queries.

Diab & Finch [6] propose a statistical word-level translation model for comparable corpora, which finds a cross-linguistic mapping between the words in the two corpora such that the source language word-level co-occurrences are preserved as closely as possible.

Finally, there is a lot of research on string similarity which has been or potentially could be applied to cognate identification: Ristad&Yianilos'98 [29] learn the MED weights using a stochastic transducer. Tiedemann'99 [32] and Mulloni&Pekar'06 [20] learn spelling changes between two languages for LCSR and for NEDR respectively. Kondrak'05 [14] proposes longest common prefix ratio, and longest common subsequence formula, which counters LCSR's preference for short words. Klementiev&Roth'06 [10] and Bergsma&Kondrak'07 [3] propose a discriminative frameworks for string similarity. Brill&Moore'00 [4] learn string-level substitutions.

6 Conclusion and Future Work

We have proposed and demonstrated the potential of a novel method for improving word alignments using linguistic knowledge and the Web as a corpus.

There are many things we plan to do in the future. First, we would like to replace competitive linking with maximum weight bipartite matching. We also want to improve MMED by adding more linguistically knowledge or by learning the NEDR or LCSR weights automatically as described in [20, 29, 32]. Even better results could be achieved with string-level substitutions [4] or a discriminative approach [3, 10] .

References

[1] Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F. Och, D. Purdy, N. Smith, and D. Yarowsky. Statistical machine translation. Technical report, 1999.

[2] N. Ayan and B. Dorr. Going beyond AER: an extensive analysis of word alignments and their impact on MT. In *Proceedings of ACL*, pages 9–16, 2006.

[3] S. Bergsma and G. Kondrak. Alignment-based discriminative string similarity. In *Proceedings of ACL*, pages 656–663, 2007.

[4] E. Brill and R. C. Moore. An improved error model for noisy channel spelling correction. In *Proceedings of ACL*, pages 286–293, 2000.

[5] P. Brown, J. Cocke, S. D. Pietra, V. D. Pietra, F. Jelinek, R. Mercer, and P. Roossin. A statistical approach to language translation. In *Proceedings of the 12th conference on Computational linguistics*, pages 71–76, 1988.

[6] M. Diab and S. Finch. A statistical word-level translation model for comparable corpora. In *Proceedings of RIAO*, 2000.

[7] P. Fung and L. Yee. An IR approach for translating from non-parallel, comparable texts. In *Proceedings of ACL*, volume 1, pages 414–420, 1998.

[8] W. Gale and K. Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102, 1993.

[9] U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada. Fast decoding and optimal decoding for machine translation. In *Proceedings of ACL*, pages 228–235, 2001.

[10] A. Klementiev and D. Roth. Named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of HLT-NAACL*, pages 82–88, June 2006.

[11] P. Koehn. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA*, pages 115–124, 2004.

[12] P. Koehn and K. Knight. Learning a translation lexicon from monolingual corpora. In *Proceedings of ACL workshop on Unsupervised Lexical Acquisition*, pages 9–16, 2002.

[13] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of NAACL*, pages 48–54, 2003.

[14] G. Kondrak. Cognates and word alignment in bitexts. In *Proceedings of the 10th Machine Translation Summit*, pages 305–312, 2005.

[15] G. Kondrak, D. Marcu, and K. Knight. Cognates can improve statistical translation models. In *Proceedings of HLT-NAACL 2003 (companion volume)*, pages 44–48, 2003.

[16] V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, (10):707–710, 1966.

[17] D. Marcu and W. Wong. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of EMNLP*, pages 133–139, 2002.

[18] D. Melamed. Automatic evaluation and uniform filter cascades for inducing N-best translation lexicons. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 184–198, 1995.

[19] D. Melamed. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, 2000.

[20] A. Mulloni and V. Pekar. Automatic detection of orthographic cues for cognate recognition. In *Proceedings of LREC*, pages 2387–2390, 2006.

[21] F. J. Och. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, 2003.

[22] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), 2003.

[23] F. J. Och and H. Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449, 2004.

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318, 2002.

[25] E. Paskaleva. Compilation and validation of morphological resources. In *Workshop on Balkan Language Resources and Tools (Balkan Conference on Informatics)*, pages 68–74, 2003.

[26] E. Paskaleva and S. Mihov. Second language acquisition from aligned corpora. In *Proceedings of Language Technology and Language Teaching*, pages 43–52, 1998.

[27] R. Rapp. Identifying word translations in non-parallel texts. In *Proceedings of ACL*, pages 320–322, 1995.

[28] R. Rapp. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of ACL*, pages 519–526, 1999.

[29] E. Ristad and P. Yianilos. Learning string-edit distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(5):522–532, 1998.

[30] M. Silberztein. *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Masson, Paris, 1993.

[31] K. Sparck-Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[32] J. Tiedemann. Automatic construction of weighted string similarity measures. In *Proceedings of EMNLP-VLC*, pages 213–219, 1999.

[33] A. Zaliznyak. *Grammatical Dictionary of Russian.* Russky yazyk, Moscow, 1977 (А. Зализняк, *Грамматический словарь русского языка.* “Русский язык”, Москва, 1977).