Computer Science Journal of Moldova, vol.**21**, no.3(63), 2013

# Wiki-Translator: Multilingual Experiments for In-Domain Translations

Dan Tufiş, Radu Ion, Ştefan Daniel Dumitrescu

**Abstract**

The benefits of using comparable corpora for improving translation quality for statistical machine translators have been already shown by various researchers. The usual approach is starting with a baseline system, trained on out-of-domain parallel corpora, followed by its adaptation to the domain in which new translations are needed. The adaptation to a new domain, especially for a narrow one, is based on data extracted from comparable corpora from the new domain or from an as close as possible one. This article reports on a slightly different approach: building an SMT system entirely from comparable data for the domain of interest. Certainly, the approach is feasible if the comparable corpora are large enough to extract SMT useful data in sufficient quantities for a reliable training. The more comparable corpora, the better the results are. Wikipedia is definitely a very good candidate for such an experiment. We report on mass experiments showing significant improvements over a baseline system built from highly similar (almost parallel) text fragments extracted from Wikipedia. The improvements, statistically significant, are related to what we call the level of translational similarity between extracted pairs of sentences. The experiments were performed for three language pairs: Spanish-English, German-English and Romanian-English, based on sentence pairs extracted from the entire dumps of Wikipedia as of December 2012. Our experiments and comparison with similar work show that adding indiscriminately more data to a training corpus is not necessarily a good thing in SMT.

**Keywords:** comparable corpora, extraction of parallel sentences, language model, statistical machine translation, translation models.

# 1 Introduction

The research on domain adaptation based on comparable corpora has been motivated by the scarce parallel data for most of the language pairs or by the scarcity of (narrow) domain specific parallel data. The standard approach is to start with a baseline system, trained on as much as possible out-of-domain parallel corpora, followed by its adaptation to the domain in which new translations are needed. To date, OPUS[1] (Tiedemann, 2012) is the largest **online** collection of parallel corpora, comprising juridical texts (EUROPARL and EUconst), medical texts (EMEA), technical texts (e.g. software KDE manuals, PHP manuals), movie subtitles corpora (e.g. OpenSubs), translated transcribed talks (e.g. TED) or news (SETIMES) but these corpora are not available for all language pairs nor their sizes are similar with respect to the domain. Another example of a large collection of aligned parallel texts (in 22 languages) is JRC-Acquis (Steinberger et al., 2006), the total body of European Union (EU) law applicable in the EU Member States.

The adaptation to a new domain, especially for a narrow one, is based on data extracted from comparable corpora from the new domain or from an as close as possible one.

This article[2] reports on slightly different approach: building an SMT system entirely from comparable data for the domain of interest. The approach is feasible if the comparable corpora are large enough to extract SMT useful data in sufficient quantities for a reliable training. If the corpora, from which the translation-useful data are searched for, are strongly comparable, the outcomes may be surprisingly good.

Wikipedia is definitely a very good candidate for such an experiment. Wikipedia is not a real parallel corpus, but a strongly comparable multilingual corpus with many documents in different languages representing translations from (mainly) English. More often than not, the documents in one language are shortened or adapted translations of

[1] http://opus.lingfil.uu.se/

[2] A preliminary version of the results have been described in (Tufiş et al, 2013); here we bring more and new experimental results and comments.

documents from other (not always the same) languages and this property of Wikipedia together with its size makes it the ideal candidate of a strongly comparable corpus from which parallel sentences can be mined.

SMT engines like Moses[3] produce better translations when presented with larger and larger parallel corpora. In this context, large and good quality parallel corpora extracted from Wikipedia for different language pairs, can serve three purposes:

1. provide in-domain training data for aiding automatic translation of English (or other languages) Wikipedia articles into other languages thus paving the way to growing for poorer foreign Wikipedia sites;

2. provide in-domain training data for aiding automatic translation of Wikipedia non-English articles thus helping the dissemination of other nations' cultural and scientific contributions;

3. add a new domain (for many language pairs), the encyclopedic domain, to the list of domains for which parallel data already exist.

The structure of the article is as follows: we begin with a short review of related research (Section 2), continue with an informal description of the tool that was used to collect the parallel sentences from Wikipedia (Section 3). In Section 4, we describe the two-steps methodology for data extraction and provide quantitative data about the obtained parallel corpora for the English-Spanish, English-German and English-Romanian language pairs. We also provide BLEU evaluation of SMT using extracted data. Next, in Section 5 we compare our experiments with similar ones. The last section draws some conclusions and presents future plans.

[3]http://www.statmt.org/moses/

## 2 Related work

Adafre and Rijke (2006) were among the first to attempt extraction of parallel sentences from Wikipedia. Their approach consists of two experiments: 1) the use of an MT system (Babelfish) to translate from English to Dutch and then, by word overlapping, to measure the similarity between the translated sentences and the original sentences and 2) with an automatically induced (phrase) translation lexicon from the titles of the linked articles, they measure the similarity of source (English) and target (Dutch) sentences by mapping them to (multiple) entries in the lexicon and computing lexicon entry overlap. Experiments were performed on 30 randomly selected English-Dutch document pairs yielding a few hundred parallel sentence pairs.

Mohammadi and GhasemAghaee (2010) continue the work of Adafre and Rijke (2006) by imposing certain limits on the sentence pairs that can be formed from a Wikipedia document pair: the length of the parallel sentence candidates must correlate and the Jaccard similarity of the lexicon entries (seen as IDs) mapped to source (Persian) and target (English) must be as high as possible. As with Adafre and Rijke, the work performed by Mohammadi and GhasemAghaee does not actually generate a parallel corpus but only a couple hundred parallel sentences intended as a proof of concept.

Another experiment, due to Smith et al. (2010), addressed large-scale parallel sentence mining from Wikipedia. They automatically extracted large volumes of parallel English-Spanish (almost 2M pairs), English-German (almost 1.7M pairs) and English-Bulgarian (more than 145K pairs) sentences using binary Maximum Entropy classifiers (Munteanu and Marcu, 2005). The work of Smith et al. (2010) is the only one we are aware of, which extracted parallel corpora of similar sizes to ours. They released their Wikipedia test sets for English-Spanish (500 pairs) and for English-German (314 pairs), an inescapable opportunity for a direct comparison between our results and theirs. This comparison is documented in the Section 5.

## 3 Extracting bilingual comparable translation units

The EU project ACCURAT[4] (2010-2013) collected from the web very large sets of comparable documents and classified them (using a specially designed metrics) into different comparability classes: strongly comparable, comparable, weakly comparable and unrelated documents (see the public site for detailed reports and the associated data). From different comparability classes, various text mining systems, developed within the project, extracted useful MT data (highly similar cross-lingual sentences and parallel terms and name entities) which subsequently were used for assessing the impact on the translation quality of the existing baseline systems (Skadia et al., 2012; Tufiş, 2012).

For the experiments described in this article we used one of the ACCURAT text miners, namely LEXACC (Ştefănescu et al., 2012). It is a fast algorithm for parallel sentence mining from comparable corpora, developed to handle large amounts of comparable corpora in a reasonable amount of time. Unlike most text-miners based on binary classifiers (e.g. Munteanu and Marcu (2005)) which do not make the distinction between truly parallel sentences, partial parallel sentences, strongly comparable or weakly comparable sentences (or other, finer degrees of parallelism), LEXACC uses a similarity metrics allowing for ranking translation candidate pairs according to their similarity scores (with values continuously ranging from a very low number assigned to unrelated sentences to a very high number assigned to truly parallel sentences).

In order to significantly reduce the search space, LEXACC uses Lucene[5] to index the entire collection of target sentences (storing the document pair ID with each sentence). Using CLIR techniques the candidate sentence pairs are subject to several restricting filters (e.g. the length of the source and target sentence candidates must correlate, a high proportion of the source sentence content words must have a translation in the target candidate, etc.). In order to extract more in-

[4]http://www.accurat-project.eu/
[5]http://lucene.apache.org/

formative sentence pairs, LEXACC filters out titles or short sentences (with less than 3 words). Once this fast initial filtering is finished, a second step, computationally more expensive, generates the final similarity ranking of the translation pairs and leaves out all the pairs with a score below a pre-established threshold.

The translation similarity measure is a weighted sum of feature functions that indicate if the source piece of text is translated by the target. Given two sentences, $s$ in the source language and $t$ in the target language, then the translation similarity measure $P(s,t)$ is:

$$P(s,t) = \sum_i \theta_i f_i(s,t) \tag{1}$$

such that $\sum_i \theta_i = 1$. Each feature function $f_i(s,t)$ returns a real value between 0 ($s$ and $t$ are not related at all) and 1 ($t$ is a translation of $s$) and contributes to the overall parallelism score with a specific fraction $\theta_i$ that is language-pair dependent and that is automatically determined by training a logistic regression classifier on existing parallel data in both directions: source-target and target-source. It follows that the translation similarity measure possible values are between 0 ($s$ and $t$ are not related at all) and 1 ($t$ is a translation of $s$).

Some of the features used by the translation similarity measure in equation 1 are as follows (for a detailed description of these features, the reader is directed to (Ştefănescu et al., 2012)):

- the content words translation strength (i.e. the score of the best alignment between content words of $s$ and $t$);
- the functional words translation strength (i.e. the score of the best alignment of functional words near a strong alignment link of content words);
- alignment obliqueness (i.e. the score of a content word alignment whose links do not cross is larger than the score of an alignment with crossing links);

- the *sentinel* translations feature: we noticed that, more often than not, two parallel sentences begin and end with strongly related content words even if words in the middle are not found in the lexicon.

In order to use LEXACC for mining useful MT sentence pairs one needs a translation lexicon. Ideally, this lexicon should be domain-specific, with a large lexical coverage of the search space. However, this requirement, difficult to meet, may be avoided using a boosting technique: use any available bilingual lexicon or extract a translation lexicon from whatever parallel corpora; run LEXACC on the in-domain comparable corpus and use the mined sentence pairs for extracting better in-domain lexicons; redo the sentence pair extraction. The boosting procedure may be repeated a number of times until no improvements are observed. Yet, one has to consider that the entire chain of processing is highly computational intensive, and depending on the size of the search space (as is the case of large Wikipedias) it may take several days.

## 4 Mining Wikipedia

Among the 285 language editions of Wikipedia (http://meta.wikimedia.org/wiki/List_of_Wikipedias[6]) created under the auspices of Wikimedia Foundation, the English, German and Spanish ones are listed in the best populated category, with more than 1,000,000 articles: English is the largest collection with 4,238,043 articles, German is the second largest with 1,587,660 articles while Spanish is the $6^{th}$ in the top Wikipedias with 1,017,938 articles. Romanian Wikipedia is in the medium populated category, and with 226,004 articles is the $25^{th}$ largest collection. For our experiments we selected three very large Wikipedias (English, German and Spanish) and a medium sized Wikipedia (Romanian) and performed SMT experiments for three language pairs: English-German, English-Spanish and English-Romanian.

[6]Consulted on May $22^{nd}$ 2013:

With these monolingual Wikipedias selected for parallel sentence mining, we downloaded (December $22^{nd}$, 2012) the "database backup dumps"[7] for which Wikipedia states that they contain "a complete copy of all Wikimedia wikis, in the form of wikitext source and metadata embedded in XML". Parsing the English XML dump, we kept only the proper encyclopedic articles which contain links to their corresponding articles in the Spanish, German or Romanian. Thus, we removed articles that were *talks* (e.g. Talk:Atlas Shrugged), *logs* (e.g. Wikipedia:Deletion log), *user related articles* (e.g. User:AnonymousCoward), *membership related articles* (e.g. Wikipedia:Building Wikipedia membership), *manuals* and *rules related articles*, etc.

For each language, the retained articles were processed using regular expressions to remove the XML mark-up in order to keep only the raw, UTF-8 encoded text, which was saved into a separate file. The non-textual entries like images or tables were stripped off. Each text document was then sentence-split using an in-house freely available[8] sentence splitter based on a Maximum Entropy classifier.

Table 1. Linked documents for three language pairs

| **Language pair** | **Document pairs** | **Size on disk** | **Size ratio (L1/L2)** |
|---|---|---|---|
| English-German | 715,555 | 2.8 Gb (English) | 1,22 |
| | | 2.3 Gb (German) | |
| English-Romanian | 122,532 | 778.1 Mb | 3,91 |
| | | 198.9 Mb | |
| English-Spanish | 573,771 | 2.5 Gb | 1,66 |
| | | 1.5 Gb | |

Table 1 lists the number of sentence-split Wikipedia comparable document pairs (by following the inter-lingual links) for each considered

[7] http://dumps.wikimedia.org/backup-index.html
[8] http://nlptools.racai.ro/nlptools/index.php?page=ssplit

language pair (see (Ştefănescu and Ion, 2013) for further details).

Looking at the size ratio of the linked documents for each language pair it is apparent that Romanian documents are much shorter than the linked English ones. The size ratios for other language pairs are more balanced, coming closer to expected language specific ratio for a parallel corpus (see below).

We applied the boosting procedure as follows:

a) We used the JRC-Acquis parallel corpora to extract initial translation lexicons for English-Romanian and English-German language pairs. For English-Spanish pair we used the corresponding parallel sub-part of EUROPARL. We run GIZA++ (Gao and Vogel, 2008) and symmetrized the extracted translation lexicons between the source and target languages. The Romanian-English lexicon extracted with GIZA++ was merged with an in-house dictionary generated from our wordnet (Tufiş et al., 2013) aligned to Princeton WordNet. With these lexicons we performed the first phase of LEXACC extraction of comparable sentence pairs from the respective Wikipedias. Let us call this data-set, for a language pair L1-L2, as Wiki-Base (L1, L2). The experiments with Wiki-Base for three language pairs are described in Section 4.2;

b) From the most MT useful parts of Wiki-Base(L1, L2), as resulted from the first step, we extracted new translation lexicons used for a second phase (the boosting) of LEXACC (symmetrized) extraction, thus getting a new and larger data set which we refer to as Wiki-Train (L1,L2). The most useful parts of Wiki-Train were identified based on their impact on the BLEU score for the test set as described in Section 4.3 and used for the training of the Wiki-Translators.

### 4.1 Building Wiki-Base (L1,L2)

Table 2 lists, for different similarity scores as extraction thresholds, the number of MT useful sentence pairs (P) found in each language pair dataset, as well as the number of words (ignoring punctuation) per

language (English Words, German Words, Romanian Words, Spanish Words) in the respective sets of sentence pairs. Obviously, data extracted with a given Similarity score threshold was a proper sub-set of any data extracted with a lower Similarity score threshold.

Table 2: Wiki-base: number of parallel sentences and words for each language pair, for a given threshold

| **Similarity score** | **English-Romanian** | **English-German** | **English-Spanish** |
|---|---|---|---|
| 0.9 | Pairs: 42,201 English Words: 0.814 M Romanian Words: 0.828 M | Pairs: 38,390 English Words: 0.554 M German Words: 0.543 M | Pairs: 91,630 English Words: 1.126 M Spanish Words: 1.158 M |
| 0.8 | Pairs: 112,341 English Words: 2.356 M Romanian Words: 2.399 M | Pairs: 119,480 English Words: 2.077 M German Words: 2.010 M | Pairs: 576,179 English Words: 10.504 M Spanish Words: 11.285 M |
| 0.7 | Pairs: 142,512 English Words: 2.987 M Romanian Words: 3.036 M | Pairs: 190,135 English Words: 3.494 M German Words: 3.371 M | Pairs: 1,219,866 English Words: 23.730 M Spanish Words: 25.931 M |
| 0.6 | Pairs: 169,662 English Words: 3.577 M Romanian Words: 3.634 M | Pairs: 255,128 English Words: 4.891 M German Words: 4.698 M | Pairs: 1,579,692 English Words: 31.022 M Spanish Words: 33.706 M |

Continuation of Table 2

| **Similarity score** | **English-Romanian** | **English-German** | **English-Spanish** |
|---|---|---|---|
| 0.5 | Pairs: 201,263 English Words: 4.262 M Romanian Words: 4.325 M | Pairs: 322,011 English Words: 6.453 M German Words: 6.186 M | Pairs: 1,838,794 English Words: 36.512 M Spanish Words: 39.545 M |
| 0.4 | Pairs: 252,203 English Words: 5.415 M Romanian Words: 5.482 M | Pairs: 412,608 English Words: 8.470 M German Words:8.132 M | Pairs: 2,102,025 English Words: 42.316 M Spanish Words: 45.565 M |
| 0.3 | Pairs: 317,238 English Words: 6.886 M Romanian Words: 6.963 M | Pairs: 559,235 English Words: 11.797 M German Words: 11.353M | Pairs: 2,656,915 English Words: 54.932 M Spanish Words: 58.524 M |

Depending on the similarity threshold, the extracted pairs of sentences may be really parallel, may contain real parallel fragments, may be similar in meaning but with a different wording, or lexically unrelated in spite of domain similarity. That is, the lower the threshold, the higher the noise.

By random manual inspection of the generated sentence pairs, we saw that, in general, irrespective of the language pair, sentence pairs with a translation similarity measure higher than 0.6 are parallel. Based on the number of words in each language side of the parallel extracted sentences, one can easily compute an expected average length ratio for the three considered language pairs. Those pairs with a translation similarity measure of at least 0.5 have extended parallel fragments which an accurate word or phrase aligner easily detects.

Further down the threshold scale, below 0.3, we usually find sentences that roughly speak of the same event but are not actual translations of each other. The noisiest data sets were extracted for the 0.1 and 0.2 similarity thresholds and we drop them from further experiments.

If we consider the extraction rate ExtR as the ratio between the number of parallel sentences (those with similarity score higher or equal to 0.7) and the number of linked documents we get the following figures:

$$\mathrm{ExtR}(\mathrm{En}-\mathrm{Ro}) = 1.16; \quad \mathrm{ExtR}(\mathrm{En}-\mathrm{De}) = 0.26;$$

$$\mathrm{ExtR}(\mathrm{En}-\mathrm{Es}) = 2.12.$$

The striking differences may have several explanations. The first one is that the Spanish documents linked to English documents are more literary translated, while the German documents are more distant from the English documents to which they are linked. Romanian documents are somewhere in between. Another partial explanation might be the quality of the dictionaries LEXACC used for each language. Augmenting the Romanian-English lexicon extracted by GIZA++ from JRC-Acquis with the data from wordnet resulted in a cleaner (although smaller) dictionary than the German-English extracted also from JRC-Acquis. In case of Spanish-English extraction rate (higher than for other two language pairs) we hypothesize that the GIZA++ dictionary extracted from EUROPARL has a better covering of the Wikipedia vocabulary. The experimental results described in the following sections strongly support these hypotheses.

### 4.2 SMT experiments with Wiki-Base

In order to select the most MT useful parts of Wiki-Base for the three considered language pairs, we built three baseline Moses-based SMT systems using only parallel sentences, that is those pairs extracted with a similarity score higher or equal to 0.7 (see Table 2). We incrementally extended the training data by lowering the similarity score threshold and, using the same test-set, observed the variation of the BLEU score. The purpose for the evaluation of the SMT systems was only to indicate what would be the best threshold for selecting the

training set from the Wiki-Train for building the Wiki-Translators. As the standard SMT system we chose Moses surface-to surface translation and lexical reordering model with parameters wbe-msd-bidirectional-fe, with phrase-length of maximum 4 words, and the default values for the rest of parameters.

The **language model** (LM) for all experiments was trained on all monolingual, sentence-split English Wikipedia after removing the administrative articles as described in Section 3. The language model was limited to 5-grams and the counts were smoothed by the interpolated Knesser-Ney method.

Since we experimentally noticed that the additional sentence pairs extracted for a threshold of 0.6 were almost as parallel as those extracted for higher thresholds we included this interval too in the sampling process for **test set** design. Thus, we proceeded to randomly sample 2,500 sentence pairs from similarity intervals ensuring parallelism ([0.6, 0.7), [0.7-0.8), [0.8, 0.9) and [0.9-1]). We obtained 10,000 parallel sentence pairs for each language pair. Additionally, we extracted 1,000 sentence pairs as development set (**dev set**). These 11,000 sentences were removed from the training corpora of each language pair. When sampling parallel sentence pairs, we were careful to obey the Moses' filtering constraints: both the source and target sentences must have at least 4 words and at most 60 words and the ratio of the longer sentence (in tokens) of the pair over the shorter one must not exceed 2. The duplicates were also removed.

Further on, we trained **seven translation models** (TM), for each language pair, over cumulative threshold intervals beginning with 0.3: $TM_1$ for [0.3, 1], $TM_2$ for [0.4, 1] ..., $TM_7$ for [0.9, 1]. The resulting eight training corpora have been filtered with Moses' cleaning script with the same restrictions mentioned above. For every language, both the training corpora and the test set have been tokenized using Moses' tokenizer script and true-cased. The quality of the translation systems is measured as usual in terms of their BLUE score (Papineni et al., 2002) on the same test data.

We have to emphasize that the removal of the sentences in the test and development sets from the training corpora does not ensure an un-

biased evaluation of the BLEU scores since their context still remained in the training corpora. This requires some more explanations. For each extracted sentence pair, LEXACC stores in a book-keeping file, the ID of the document-pair out of which the extraction was done. This information allows for elimination from the training set of all the pairs coming from the same documents from which the development and evaluation sets were selected. However, due to the nature of the Wikipedia article authoring, this strategy of filtering the development and evaluation sets does not ensure an unbiased evaluation. The Wikipedia contributors are given specific instructions for authoring documents[9] and by observing these instructions, inherently one could find in different documents almost identical sentences except for a few name entities. Indeed we found examples of such sentence pairs in the train set similar, but not identical, to sentences in the test set, yet coming from different document-pairs. Certainly one could build a tough test-set by removing from train set all similar (pattern-based) sentences, but we did not do that because it would have been beyond the purpose of this work. As we mentioned before, this evaluation was meant only for estimating most useful extraction level for the second phase of training the WIKI-Translators.

Table 3 summarizes the results of this first step experiment, the bold characters identifying the most MT useful parts of Wiki-Base (L1,L2). We considered TM $_{[0.7,1]}$ (the shaded line in Table 3) as the baseline for all language pairs.

### 4.3 Building Wiki-Train (L1,L2)

The experiments on Wiki-base revealed that the most useful training data has been extracted by using LEXACC with 0.5 similarity score for German-English and Romanian-English language pairs and 0.3 for Spanish-English pair (see Table 3). We re-run GIZA++ on these subsets of Wiki-Base to extract new in-domain lexicons.

The new lexicons were merged with the initial ones and the LEXACC extraction was repeated with the resulted mined comparable

[9] http://en.wikipedia.org/wiki/Wikipedia:Translation

Table 3. Comparison between SMT systems trained on various parts of Wiki-Base

| **TM based on Wiki-Base** | **BLEU SCORE Romanian –> English** | **BLEU SCORE German –> English** | **BLEU SCORE Spanish –> English** |
|---|---|---|---|
| $TM_{[0.3,1]}$ | 37.24 | 39.16 | **47.59** |
| $TM_{[0.4,1]}$ | 37.71 | 39.46 | 47.52 |
| $TM_{[0.5,1]}$ | **37.99** | **39.52** | 47.53 |
| $TM_{[0.6,1]}$ | 37.85 | 39.5 | 47.44 |
| $TM_{[0.7,1]}$ | 37.39 | 39.24 | 47.28 |
| $TM_{[0.8,1]}$ | 36.89 | 38.57 | 46.27 |
| $TM_{[0.9,1]}$ | 32.76 | 34.73 | 39.68 |

sentence-pairs denoted as Wiki-Train.

As the experiments on the Wiki-base showed that for a similarity threshold less than or equal to 0.2 LEXACC delivers not very useful data, we started the second step of mining using the similarity scores of at least 0.3.

Table 4 shows the results of the boosted extraction process. As one can see the extracted data, at each similarity score level, is significantly increased for the English-Romanian and English-German language pairs. For English-Spanish, except for the similarity scores 0.8 and 0.9 the number of sentence pairs is smaller than in Wiki-Base. The reason is that in this round we detected several identical pairs with those in the training and development sets and several duplicated pairs in the training set. Anyway, the English-Spanish Wiki-Train was the largest train-set and containing the highest percentage of fully parallel sentence pairs.

Table 4: Wiki-Train: number of parallel sentences and words for each language pair, for a given threshold

| **Similarity score** | **English-Romanian** | **English-German** | **English-Spanish** |
|---|---|---|---|
| 0.9 | Pairs: 66,777 English Words: 1.077 M Romanian Words: 1.085 M | Pairs: 97,930 English Words: 1.069 M German Words: 1.042 M | Pairs: 113,946 English Words: 1.164 M Spanish Words: 1.193 M |
| 0.8 | Pairs: 152,015 English Words: 2.688 M Romanian Words: 2.698 M | Pairs: 272,358 English Words: 3.695M German Words: 3.552 M | Pairs: 597,992 English Words: 9.733 M Spanish Words: 10.510 M |
| 0.7 | Pairs: 189,875 English Words: 3.364 M Romanian Words: 3.372 M | Pairs: 434,019 English Words: 6.201 M German Words: 5,929 M | Pairs: 1,122,379 English Words: 19.941 M Spanish Words: 21.821 M |
| 0.6 | Pairs: 221,661 English Words: 3.961 M Romanian Words: 3.970 M | Pairs: 611,868 English Words: 8.944 M German Words8.532 M | Pairs: 1,393,444 English Words: 25.068 M Spanish Words: 27.411 M |
| 0.5 | Pairs: 260,287 English Words: 4,715 M Romanian Words: 4,722 M | Pairs: 814,041 English Words: 12.361 M German Words: 11.792M | Pairs: 1,587,276 English Words: 28.987 M Spanish Words: 31.567 M |

Continuation of Table 4

| **Similarity score** | **English-Romanian** | **English-German** | **English-Spanish** |
|---|---|---|---|
| 0.4 | Pairs: 335,615 English Words: 6.329 M Romanian Words: 6.324 M | Pairs: 1,136,734 English Words: 18,089 M German Words: 17.306 M | Pairs: 1,807,892 English Words: 33.619 M Spanish Words: 36,369 M |
| 0.3 | Pairs: 444,102 English Words: 8.712 M Romanian Words: 8.700 M | Pairs: 1,848,651 English Words: 31.405 M German Words: 30.175 M | Pairs: 2,288,163 English Words: 44.021 M Spanish Words: 47.180 M |

### 4.4 SMT experiments with Wiki-Train

The Wiki-Train corpora were used with the same experimental setup as described in Section 4.2. The training of each translation system was followed by the evaluation on the respective test sets (10,000 pairs) in both translation directions. To make the comparison between the translation qualities we did the translations without MERT optimization of the parameters. The results are presented in Table 5.

Having much more training data, in case of the Romanian –> English and German –>English the BLEU scores significantly increased (with 3.1 and 2.58 points respectively). For Spanish-English the decrease of number of sentences in Wiki-Train as compared to Wiki-Base negatively impacted the new BLEU score, which is 1.31 points lower. It would be interesting to see what would happen with a higher threshold training set, for instance $TM_{[0.5,1]}$, as used for the other language pairs.

As expected, the translations into non-English languages are less accurate due to a more complex morphology of the target language (most of the errors are morphological ones), but still the BLEU scores

are very high, better than most of the results we are aware off (for in-domain experiments).

Table 5. Best translation SMT systems, trained on Wiki-Train[10]

| **TM based on Wiki-Train** | **$TM_{[0.5,1]}$ Romanian –> English** | **$TM_{[0.5,1]}$ German –> English** | **$TM_{[0.3,1]}$ Spanish –> English** |
|---|---|---|---|
| **BLEU SCORE** | **41.09** | **40.82** | **46.28** |
| | **$TM_{[0.5,1]}$ English –> Romanian** | **$TM_{[0.5,1]}$ English –> German** | **$TM_{[0.3,1]}$ English –> Spanish** |
| **BLEU SCORE** | **29.61** | **35.18** | **46.00** |

## 5 Comparison with other works

Translation for Romanian-English language pair has also been studied in (Boroş et al., 2013; Dumitrescu et al., 2012; 2013) among others. In these works we had explicit interests in experiments on using in-domain/out-of-domain test/train data, and various configurations of the Moses decoder in surface-to-surface and factored translation. Out of the seven domain-specific corpora (Boroş et al., 2013) one was based on Wikipedia. The translation experiments on English-Romanian, similar to those reported here, were surface based (t0-0, m0) with training on parallel sentence pairs extracted from Wikipedia by LEXACC at a fixed threshold: 0.5 (called "WIKI5"), without MERT optimization. A random selection of unseen 1,000 Wikipedia Romanian test sentences[11] has been translated into English using combinations of:

- a WIKI5-based translation model (240K sentence pairs)/WIKI5-based language model;

[10]For a fair comparison with data in Table 3 we did not use here the MERT optimization

[11]The test-set construction followed the same methodology described in this article

- a global translation model (1.7M sentence pairs)/global language model named "ALL", made by concatenating all specific corpora.

Table 6 gives the BLEU scores for the Moses configuration similar to ours.

Table 6. BLEU scores on 1000 sentences Wikipedia test set of Dumitrescu et al. (2013)

| | **WIKI5 TM** | **ALL TM** |
|---|---|---|
| **WIKI5 LM** | **29.99** | 29.95 |
| **ALL LM** | 29.51 | 29.95 |

Boroş et al.'s results confirm the conclusion we claimed earlier: the ALL system performs worse than the in-domain WIKI5 system. The large difference between the herein BLEU score (41.09) and 29.99 in (Boroş et al., 2013) may be explained by various factors. First and more importantly, our current language model was entirely in-domain for the test data and much larger: the language model was built from entire Romanian Wikipedia (more than 220,000 documents) while the language model in (Boroş et al., 2013) was built only from the Romanian sentences paired to English sentences (less than 240,000 sentences). Our translation model was built from more than 260,000 sentence pairs versus 234,879 sentence pairs of WIKI5). Another explanation might be the use of different Moses filtering parameters (e.g. the length filtering parameters) and different test sets. As suggested by other researchers, Wikipedia-like documents are more difficult to translate than, for instance, legal texts. The BLEU scores on JRC-Acquis test sets (with domain specific training) reported in (Boroş et al., 2013) is almost double than those obtained on Wikipedia test sets.

The most similar experiments to ours have been reported by Smith et al. (2010). They mined for parallel sentences from Wikipedia producing parallel corpora of sizes even larger than ours. While they used for training all the extracted sentence pairs, we used only those subsets that observed a minimal similarity score. We checked to see if their test sets for English-Spanish (500 pairs) and for English-German

(314 pairs) contained sentences in our training sets and, as this was the case, we eliminated from the training several sentence pairs (about 200 sentence pairs from the English-Spanish training corpus and about 140 sentence pairs from the English-German training corpus). We re-trained the two systems on the slightly modified training corpora. Since in their experiments they used MERT-optimized translation systems, we optimized, also by MERT, our new $TM_{[0.5,1]}$ for German–>English and new $TM_{[0.3,1]}$ for Spanish–>English translation systems, using the respective dev-sets (each containing 1,000 sentence pairs).

Their test sets for English-Spanish and for English-German were translated (after being true-cased) with our best translation models and also with Google Translate (as of mid-February 2013).

Table 7 summarizes the results. In this table, "Large+Wiki" denotes the best translation model of Smith et al. which was trained on many corpora (including Europarl and JRC Acquis) and on more than 1.5M parallel sentences mined from Wikipedia. "$TM_{[0.4,1]}$" and "$TM_{[0.5,1]}$" are our Wiki-Train translation models as already explained. "Train data size" gives the size of training corpora in multiples of 1,000 sentence pairs.

Table 7. Comparison between SMT systems on the Wikipedia test set provided by Smith et al. (2010)

| **Language pair** | **Train data size (sentence pairs)** | **System** | **BLEU** |
|---|---|---|---|
| Spanish-English | 9,642K | Large+Wiki | 43.30 |
| | 2,288K | $TM_{[0.4,1]}$ | **50.19** |
| | – | Google | 44.43 |
| German-English | 8,388K | Large+Wiki | 23.30 |
| | 814K | $TM_{[0.5,1]}$ | 24.64 |
| | – | Google | 21.64 |

For Spanish-English test set of Smith et al. (2010) our result is significantly better than theirs, in spite of almost 4 times less training

data. For the German-English pair, the difference is larger between $TM_{[0.5,1]}$ and Large+Wiki systems, and one should also notice that our system used 10 times less training data (but, presumably, much cleaner).

However, our $TM_{[0.5,1]}$ for German-English performed on the new test set much worse than on our test-set (24.64 versus 40.82[12] BLEU points) which was not the case for the Spanish-English language pair. We suspected that some German-English translation pairs in the Smith et al. (2010) test set were not entirely parallel. This idea was supported by the correlation of the evaluation results between our translations and Google's for Spanish-English and German-English. Also, their reported results on German-English were almost half of the ones they obtained for Spanish-English.

Therefore, we checked the German-English and Spanish-English test sets (supposed to be parallel) by running the LEXACC miner to see the similarity scores for the paired sentences. The results confirmed our guess. The first observation was that the test sets contained pieces of texts that looked like section titles (e.g. BT: Contaminación – BT: Pollution; Segunda clase – Second class; Autoengano-Self-deception, in Spanish – English test-set or Städte – Cities and towns; 1956 Armagnac – 1956 Armagnac; Produkte – Products; Geschwindigkeitsrekorde – Speed records; Geschichte – History in the German-English test-set). Such short sentences were ignored by LEXACC. While out of the considered sentence pairs (ignoring the sentences with less than 3 words), for Spanish-English LEXACC identified more than 92% as potentially useful SMT pairs (with a similarity score higher than or equal to 0.3 – this was the extraction threshold for Spanish-English sentence-pairs), for German-English LEXACC identified only 35% potentially useful SMT pairs (a similarity score higher than or equal to 0.5 – this was the extraction threshold for German-English sentence-pairs). Even if the threshold for German-English was lowered to 0.3, only 45% passed the LEXACC filtering. As for parallelism status of the sentence pairs in the

[12]Note that this value for our $TM_{[0.5,1]}$ was obtained on a very different and much larger test set and also without MERT optimization. Yet, the difference is large enough to raise suspicions on the test set used for this comparison.

test-sets (i.e. similarity scores higher than 0.6 for both languages) the percentages were 78% for Spanish-English and only 29% for German-English. Without ignoring the short sentences (easy to translate) these percentages would have probably been a little bit higher (80.8% for Spanish-English and 32.82% for German-English).

These evaluations outline also that LEXACC is too conservative in its rankings: we noticed almost parallel sentences in the test-set for Spanish-English even for a similarity score of 0.1 while in the German-English the same happens for similarity scores lower than 0.3. The most plausible explanation was that one of the LEXACC's parameters (cross-linking factor) strongly discourages long-distance reordering (which was quite frequent in the German-English test set and has also a few instances in the Spanish-English test set).

Table 8 shows some examples of sentence pairs in the German-English and Spanish-English test sets showing low level of parallelism (inappropriate for translation quality evaluation) but also some examples of sentence pairs which were conservatively lower ranked by LEXACC.

Table 8: Examples of sentence-pairs in the German-English and Spanish-English test sets used by Smith et al. (2010)

| **Similarity** | **German source sentence** | **English reference translation** |
|---|---|---|
| **1** | **2** | **3** |
| $< 0.1$ | Zuletzt stand sie für Robert Dornhelms Historienfilm Kronprinz Rudolf als Mary von Vetsera und in Le Ragazze di San Frediano als Mafalda vor der Filmkamera. | Puccini 's role as Mafalda in the 2007 Rai Uno miniseries Le ragazze di San Frediano cast her among many other well-known Italian actresses, including Martina Stella, Chiara Conti, and Camilla Filippi. |

Continuation of Table 8

| **1** | **2** | **3** |
|---|---|---|
| $< 0.1$ | Unter anderem ist es für die Durchführung der Volkszählung zuständig. | Every 10 years, this organisation conducts a national census. |
| $< 0.1$ | Daraufhin nahm sich Nikolaus, der es mit der ehelichen Treue schon mehrfach nicht so genau genommen hatte, eine Mätresse, Alexandras Hofdame Barbara Nelidowa. | Nelidova went with them, and though Alexandra was jealous in the beginning, she soon came to accept the affair, and remained on good terms with her husband's mistress. |
| 0.12 | Im Unterschied zu Cognac wird Armagnac in einem kontinuierlichen Brennverfahren nur einmal destilliert, also nicht rektifiziert. | Armagnac is traditionally distilled once, which results initially in a less polished spirit than Cognac, where double distillation usually takes place. |
| 0.29 | Die 64,5 Prozent , welche die SPD unter seiner Führung erzielte, waren das höchste Ergebnis, welches je eine Partei auf Bundeslandsebene bei einer freien Wahl in Deutschland erzielt hatte. | In the election that was conducted in the western part of Berlin two months later, his popularity gave the SPD the highest win with 64.5 % ever achieved by any party in a free election in Germany. |
| **Similarity** | **Spanish source sentence** | **English reference translation** |
| Misaligned | En febrero de 1988, a 12 UA del Sol, el brillo de Quirón alcanzó el 75 % Este comportamiento es típico de los cometas pero no de los asteroides. | In February 1988, 12 AU from the Sun, Chiron brightness reached 75%. |

Continuation of Table 8

| 1 | 2 | 3 |
|---|---|---|
| 0.1 | Sin embargo, el museo, llamado no fue terminado sino hasta el 10 de abril de 1981, dos días antes del vigésimo aniversario del vuelo de Yuri Gagarin. | However, it took until April 10, 1981 (two days before the 20th anniversary of Yuri Gagarin's flight) to complete the preparatory work and open the Memorial Museum of Cosmonautics. |

## 6 Conclusions

Wikipedia is a rich resource for parallel sentence mining in SMT. Comparing different translation models containing MT useful data ranging from strongly comparable, to parallel, we concluded that there is sufficient empirical evidence not to dismiss sentence pairs that are not fully parallel on the suspicion that because of the inherent noise they might be detrimental to the translation quality. On the contrary, our experiments demonstrated that in-domain comparable data are strongly preferable to out-of-domain parallel data. However, there is an optimum level of similarity between the comparable sentences, which according to our similarity metrics (for the language pairs we worked with) is around 0.4 or 0.5.

Additionally, the two step procedure we presented, demonstrated that an initial in-domain translation dictionary is not necessary, it can be constructed subsequently, starting with a dictionary extracted from whatever out-of-domain data.

We want to mention that it is not the case that our extracted Wikipedia data is the maximally MT useful data. First of all, LEXACC may be improved in many ways, which is a matter for future developments. For instance, although the cross-linking feature is highly relevant for language pairs with similar word ordering, it is not very effective for language pairs showing long distance re-ordering. We also noticed that a candidate pair for which its two parts contained different numerical entities (numbers, dates, times) was dropped from further

consideration. Thirdly, the extraction parameters of LEXACC were not re-estimated for the Wiki-Train construction. Additionally, we have to mention that LEXACC evaluated and extracted only full sentences: a finer-grained (sub-sentential) extractor would likely generate more MT useful data. Also, one should note that the evaluation figures are just indicative for the potential of Wikipedia as a source for SMT training. In previous work it was shown that using factored models for inflectional target languages (Boroş et al, 2013) and cascading translators (Tufiş and Dumitrescu, 2012) may significantly improve (several BLEU points) the translation accuracy of an SMT system. Some other techniques may be used to improve at least translations into English. For instance, given that English adjectives and all functional words are not inflected, a very effective way, for a source inflectional language would be to lemmatize all words in these categories. Another idea is to split compound words of a source language (such as German) into their constituents. Both such simplifications are, computationally, not very expensive (and for many languages appropriate tools are publicly available) but may significantly reduce the number of out-of-vocabulary input tokens.

The parallel Wiki corpora (before and after the boosting step), including the test sets (containing 10,000) and the dev–sets (containing 1,000 sentences) are freely available on-line[13].

**Acknowledgments**. This work has been supported by the EU under the Grant Agreements no. 248347 (ACCURAT) and no. 270893 (METANET4U).

## References

[1] Sisay Fissaha Adafre and Maarten de Rijke. 2006. *Finding similar sentences across multiple languages in Wikipedia.* In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), April 3-7, 2006. Trento, Italy, pp. 62–69.

[13] http://ws.racai.ro:9191/repository/search/

[2] Tiberiu Boroş, Stefan Daniel Dumitrescu, Radu Ion, Dan Ştefănescu, Dan Tufiş. 2013. *Romanian-English Statistical Translation at RACAI.* În E. Mitocariu, M. A. Moruz, D. Cristea, D. Tufiş, M. Clim (eds.) Proceedings of the 9th International Conference "Linguistic Resources and Tools for Processing the Romanian Language", 16-17 mai, 2013, Miclăuşeni, Romania, 2013. "Alexandru Ioan Cuza" University Publishing House. 197 p. ISSN 1843-911X. pp. 81–98, 2013.

[3] Ştefan Dumitrescu, Radu Ion, Dan Ştefănescu, Tiberiu Boroş, Dan Tufiş. 2013. *Experiments on Language and Translation Models Adaptation for Statistical Machine Translation.* In Dan Tufiş, Vasile Rus, Corina Forăscu (eds.) *Towards Multilingual Europe 2020: A Romanian Perspective,* pp. 205–224, 2013.

[4] Ştefan Dumitrescu, Radu Ion, Dan Ştefănescu, Tiberiu Boroş, Dan Tufiş. *Romanian to English Automatic MT Experiments at IWSLT12.* In Proceedings of the International Workshop on Spoken Language Translation, December 6 and 7, 2012, Hong Kong, pp. 136–143.

[5] Qin Gao and Stephan Vogel. 2008. *Parallel implementations of a word alignment tool.* In Proceedings of ACL-08 HLT: Software Engineering, Testing, and Quality Assurance for Natural Language Processing, June 20, 2008. The Ohio State University, Columbus, Ohio, USA, pp. 49–57.

[6] Philipp Koehn. 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation,* In Proceedings of the tenth Machine Translation Summit, Phuket, Thailand, pp. 79–86.

[7] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondej Bojar, Alexandra Constantin, and Evan Herbst. 2007. *Moses: open source toolkit for statistical machine translation.* In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and

Demonstration Sessions (ACL '07). Association for Computational Linguistics, Stroudsburg, PA, USA, pp.177–180.

[8] Mehdi Mohammadi and Nasser GhasemAghaee. 2010. *Building bilingual parallel corpora based on Wikipedia.* In Computer Engineering and Applications (ICCEA 2010), Second International Conference on Computer Engineering and Applications, Vol. 2, pp. 264–268. IEEE Computer Society Washington, DC, USA.

[9] Dragoş Munteanu, Daniel Marcu. 2005. *Improving machine translation performance by exploiting comparable corpora.* Computational Linguistics, 31(4), pp. 477–504.

[10] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. 2002. *BLEU: A method for automatic evaluation of machine translation.* In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), July 2002. Philadelphia, USA, pp. 311–318.

[11] Inguna Skadiņa, Ahmet Aker, Nikos Glaros, Fangzhong Su, Dan Tufiş, Mateja Verlic, Andrejs Vasiljevs, Bogdan Babych, 2012. *Collecting and Using Comparable Corpora for Statistical Machine Translation.* In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), May 23-26, 2012. Istanbul, Turkey, ISBN 978-2-9517408-7-7.

[12] Jason R. Smith, Chris Quirk, Kristina Toutanova. 2010. *Extracting parallel sentences from comparable corpora using document level alignment.* In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 403–411. © Association for Computational Linguistics (2010).

[13] Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş. 2006. *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages.* In Proceedings of the $5^{th}$ LREC Conference, Genoa, Italy, 22-28 May, 2006, ISBN 2-9517408-2-4, EAN 978-2-9517408-2-2.

[14] Dan Ştefănescu, Radu Ion, Sabine Hunsicker. 2012. *Hybrid parallel sentence mining from comparable corpora.* In Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT 2012), pp. 137–144, Trento, Italy, May 28-30, 2012.

[15] Jörg Tiedemann. 2012. *Parallel data, tools and interfaces in OPUS.* In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), May 23-26, 2012. Istanbul, Turkey, ISBN 978-2-9517408-7-7.

[16] Dan Tufiş. *Finding Translation Examples for Under-Resourced Language Pairs or for Narrow Domains; the Case for Machine Translation.* 2012. In Computer Science Journal of Moldova, Academy of Sciences of Moldova, Institute of Mathematics and Computer Science, ISSN 1561-4042, vol.20, no.2(59), pp. 227–245.

[17] Dan Tufiş, Verginica Barbu Mititelu, Dan Ştefănescu, Radu Ion. 2013. *The Romanian Wordnet in a Nutshell.* Language and Evaluation, Springer, Vol. 47, no. 2, 2013, ISSN 1574-020X, DOI: 10.1007/s10579-013-9230-7

[18] Dan Tufiş, Radu Ion, Ştefan Dumitrescu, Dan Ştefănescu. 2013. *Wikipedia as an SMT Training Corpus.* In Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP 2013), Hissar, Bulgaria, September 7-13, 2013.

[19] Dan Ştefănescu, Radu Ion. 2013. *Parallel-Wiki: A Collection of Parallel Sentences Extracted from Wikipedia.* In Proceedings of 14th International Conference on Intelligent Text Processing and Computational Linguistics, pp. 137–144, Trento, Italy, March 24-30, 2013, Samos, Greece.

Dan Tufiş[1], Radu Ion[2], Ştefan Daniel Dumitrescu[3] Received September 16, 2013

Institute for AI, Romanian Academy, Bucharest, Romania
[1] E-mail: tufis@racai.ro
[2] E-mail: radu@racai.ro
[3] E-mail: danstef@racai.ro