Large SMT data-sets extracted from Wikipedia

Dan Tufiş[1], Radu Ion[1], Ştefan Dumitrescu[1], Dan Ştefănescu[2]
[1]Research Institute for Artificial Intelligence - Romanian Academy,
13 Calea 13 Septembrie, 050711, Bucharest, Romania
{tufis, radu, sdumitrescu}@racai.ro
[2]The University of Memphis, Department of Computer Science,
Memphis, TN, 38152, USA-USA
dstfnscu@memphis.edu

Abstract

The article presents experiments on mining Wikipedia for extracting SMT useful sentence pairs in three language pairs. Each extracted sentence pair is associated with a cross-lingual lexical similarity score based on which, several evaluations have been conducted to estimate the similarity thresholds which allow the extraction of the most useful data for training three-language pairs SMT systems. The experiments showed that for a similarity score higher than 0.7 all sentence pairs in the three language pairs were fully parallel. However, including in the training sets less parallel sentence pairs (that is with a lower similarity score) showed significant improvements in the translation quality (BLEU-based evaluations). The optimized SMT systems were evaluated on unseen test-sets also extracted from Wikipedia. As one of the main goals of our work was to help Wikipedia contributors to translate (with as little post editing as possible) new articles from major languages into less resourced languages and vice-versa, we call this type of translation experiments "*in-genre"* translation. As in the case of "in-domain" translation, our evaluations showed that using only "in-genre" training data for translating same genre new texts is better than mixing the training data with "out-of-genre" (even) parallel texts.

Keywords: comparable corpora, in-genre translation, similarity-based text mining, Wikipedia

1. Introduction

SMT engines like Moses[1] produce better translations when presented with larger and larger parallel corpora. For a given test-set, it is also known that Moses produces better translations when presented with in-domain training data (data sampled form the same domain as the test data, e.g. news, laws, medicine, etc.), but collecting parallel data from a given domain, in sufficiently large quantities to be of use for statistical translation, is not an easy task. To date, OPUS[2] (Tiedemann, 2012) is the largest **online** collection of parallel corpora, comprising of juridical texts (EUROPARL and EUconst)[3], medical texts (EMEA), technical texts (e.g. software KDE manuals, PHP manuals), movie subtitles corpora (e.g. OpenSubs) or news (SETIMES), but these corpora are not available for all language pairs nor are their sizes similar with respect to the domain. To alleviate this data scarcity, significant research and development efforts have been invested in parallel texts harvesting from the web (Resnik & Smith, 2003) regarded as a huge multilingual comparable corpus with presumably large quantities of parallel data.

Comparable corpora have been widely recognized as valuable resources for multilingual information extraction, but few large datasets were publicly released and even fewer evaluated in the context of specific cross-lingual applications. One of the most tempting uses of comparable corpora mining is in statistical machine translation for under-resourced language pairs (e.g. Lithuanian-English) or limited/specialized domains (e.g. automotive).

Within the ACCURAT European project we developed LEXACC, a language independent text-miner for comparable corpora (Ştefănescu et al., 2012), able to extract cross-lingually similar fragments of texts, similarity being judged by means of a linear weighted combination of multiple features. The text-miner modules (sentences indexing, searching, filtering), the features and the learning of the optimal weights used by the similarity scoring function which endorses the extracted pairs are largely described in (Ştefănescu & Ion, 2013).

Unlike most comparable corpora miners that use binary classifiers, our miner associates each extracted pair (*s_t*) a symmetric similarity score, sim_score∈[0, 1]. Based on these scores, one can experiment with various subsets of an extracted data set (DS), selecting only the sentence pairs with similarity scores higher or equal to a decided threshold value. Starting with a minimal threshold value (th) for the similarity score (th_1=0.1), the experimenter will get the full data set(DS^{th_1}) which, certainly, contains many noisy pairs. Increasing the th, one gets less sentence-pairs, but more (cross-lingually) similar:$DS^{th_1} \supset DS^{th_2} \ldots \supset DS^{th_n}$, with $0.1 < \ldots th_i \ldots < 1.0$ and $DS^{th_i} = \{s\_t | \, sim\_score(s\_t) \geq th_i\}$.

The parallel text miner was used to extract sentence-pairs from large corpora collected from the web (Skadiņa et al., 2012) for a wide variety of language pairs: English-German, English-Greek, English-Estonian, English-Lithuanian, English-Latvian, English-Romanian and English-Slovene. Although the "optimal" similarity scores slightly differed from one language pair to another, they were comparable, in the interval [0.35, 0.5]. To evaluate the value of the mined parallel sentences, the extracted German-English sentences were used for domain adaptation of a baseline SMT system (trained on Europarl+news corpus). This experiment, described in (Ştefănescu et al., 2012), showed significant quality improvements over the baseline (+6.63 BLEU points) when translating texts in the automotive domain.

The evaluation of LEXACC in the ACCURAT project encouraged us to explore its efficacy on a well-known

[1] http://www.statmt.org/moses/
[2] http://opus.lingfil.uu.se/
[3] JRC-Acquis and DGT Translation Memories are other examples of large parallel juridical texts.

multilingual strongly comparable corpus, namely Wikipedia[4], and to investigate the feasibility of what we called "*in-genre*" statistical machine translation: translating documents of the same genre (not necessary the same domain) based on training data of the respective genre. Wikipedia is a strongly comparable multilingual corpus and it is large enough to ensure reasonable volumes of training data for "in-genre" translations. One useful property of Wikipedia is that, in spite of covering many domains, articles are characterized by the same writing style and formatting conventions[5]. The authors are given precise guidelines content-wise, on the language use, on formatting, and several other editorial instructions to ensure genre unity across various topics as in the traditional encyclopedias. Therefore, we refer to the translation of a Wikipedia-like article, irrespective of its topic, based on training data extracted from Wikipedia, as an "in-genre" (encyclopedic) translation. The parallel data mining processing flow (Ştefănescu & Ion, 2012) was recently improved by what we called "boosted mining" (Tufiş et al., 2013a,b), and a larger and quite different experiment was conducted with the aim of evaluating the usability of texts mined from comparable corpora for statistical translation of texts similar in genre with the extraction corpora.

2. Related work

The interlingual linking in Wikipedia makes it possible to extract documents in different languages such that the linked documents are versions of each other. If the linked documents were always faithful translations of some hub (original) documents, Wikipedia would be a huge multilingual parallel corpus of encyclopedic texts. Yet, this is not the case and, frequently, articles originally written in one language (English, most of the times) are adapted translations (usually, shortened) in other languages. It also happens that there exist articles with no links to other languages or only with links to one or two (major) languages [6]. Thus, Wikipedia is arguably the largest strongly comparable corpus available online and the best data collection for "in genre" translation experiments.

The major motivation of the current endeavour is the validation of the idea that it is possible to develop systems capable of supporting the translation (with minimal post-editing) of Wikipedia articles from less-resourced languages into major languages and vice-versa. Currently, Wikipedia does not offer an integrated translation engine to assist the translation task, but this could be an option worthy of consideration.

It is, therefore, not surprising that several researchers were tempted to explore parallel sentence mining on Wikipedia. Among the first, Adafre and Rijke (2006) approached Wikipedia mining for parallel data using a 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. They also experimented with an automatically induced (phrase) translation lexicon from the titles of the linked articles, measuring the similarity of source (English) and target (Dutch) sentences. Experiments were performed on 30 randomly selected English-Dutch document pairs yielding a few hundred parallel sentence pairs.

A few years later, Mohammadi and GhasemAghaee (2010), for another language pair (Persian-English) developed the Adafre and Rijke's method by imposing additional conditions on the extracted sentence pairs: the length of the parallel sentence candidates had to correlate and the Jaccard similarity of the lexicon entries mapped to source (Persian) and target (English) had to be as high as possible. The experiments conducted by Mohammadi and GhasemAghaee did not generate a parallel corpus, but only a couple of hundred parallel sentences intended as a proof of concept.

Gamalo and Lopez (2010) collected the linked documents in Wikipedia (CorpusPedia), building a large comparable corpus (English-Spanish, English-Portuguese and Spanish-Portuguese), with similar documents classified according to their main topics. However, they did not measure the similarity among the aligned documents neither did they extract the "parallel" sentences.

Another experiment, due to Smith et al. (2010), addressed large-scale parallel sentence mining from Wikipedia. Based on binary Maximum Entropy classifiers (Munteanu & Marcu, 2005), they automatically extracted large volumes of parallel sentences for English-Spanish (almost 2M pairs), English-German (almost 1.7M pairs) and English-Bulgarian (more than 145K pairs). According to Munteanu and Marcu (2005), a binary classifier can be trained to differentiate between parallel sentences and non-parallel sentences using various features such as: word alignment log probability, number of aligned/unaligned words, longest sequence of aligned words, etc. To enrich the feature set, Smith et al. (2010) proposed to automatically extract a bilingual dictionary from the Wikipedia document pairs and use this dictionary to supplement the word alignment lexicon derived from existing parallel corpora. Furthermore, they released their English-Spanish and English-German Wikipedia test-sets (but not the training data) and so, a direct comparison was made possible (Section 4).

3. Mining Wikipedia

Among the 287 language editions of Wikipedia (http://meta.wikimedia.org/wiki/List_of_Wikipedias [7]), 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 4,468,164 articles, German is the third largest with 1,765,430 articles (it was the second when we downloaded it), while Spanish is the 8th in the top Wikipedias with 1,086,871 articles (it was the sixth before). Romanian Wikipedia is in the medium populated category, and with 241,628 articles is the 28th largest collection (it was the 25th at the time of dump downloading). For our experiments we selected three very large Wikipedias (English, German and Spanish) and a medium-sized Wikipedia (Romanian) intended for the

[4] http://www.wikipedia.org

[5] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style, http://en.wikipedia.org/wiki/Wikipedia:TONE

[6] For instance, for Romanian-Slovak language pair, we identified only 349 linked pairs, in spite of both languages being in the category "+100,000" with more than 400,000 articles together.

[7] Consulted on March 9th 2014.The statistics on Wikipedia naturally varies at different times. All our experiments are based on the dump of December 22nd, 2012.

SMT experiments on three language pairs: English – German, English - Spanish and English - Romanian.

With the monolingual Wikipedias selected for parallel sentence mining, we downloaded the "database backup dumps"[8] 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 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 sentence splitter based on a Maximum Entropy classifier.

Language pair	Document pairs	Size on disk	Size ratio (L1/L2)
English-German	715,555	2.8 Gb (EN)	1,22
		2.3 Gb (DE)	
English-Romanian	122,532	0.78 Gb (EN)	3,91
		0.2 Gb (RO)	
English-Spanish	573,771	2.5 Gb (EN)	1,66
		1.5 Gb (ES)	

Table 1: Linked documents for three language pairs

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

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.

For extracting *SMT useful sentence pairs*[9], we applied a two steps procedure as follows:

a) We extracted the initial translation lexicons for English-Romanian and English-German language pairs from the JRC-Acquis parallel corpora. For English-Spanish pair we used the EUROPARL parallel corpus (as Smith et al. (2010) did). We ran GIZA++ (Gao & 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., 2013c) 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 SMT experiments with Wiki-Base for three language pairs are described in Section 3.2; the subsets of Wiki-Base (L1, L2) which maximized the BLEU scores were considered the most useful part for the second phase (boosting) of the experiments.

b) GIZA++ was run again and results symmetrized on the most SMT useful sentence pairs of Wiki-Base (L1, L2), as resulted from the first step. The new translation lexicons were merged with the initial ones and used for a second phase of LEXACC 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 3.3 and used for the training of the Wiki-Translators.

3.1. Phase 1: Building Wiki-Base (L1, L2)

Table 2 lists, for different similarity scores as extraction thresholds, the number of SMT 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. As mentioned before, data extracted with a given Similarity score threshold is a proper subset of any data extracted with a lower Similarity score threshold.

Sim. score	EN-RO	EN-DE	EN-ES
0.9	Pairs: 42,201 EN Words: 0.81M RO Words: 0.83M	Pairs: 38,390 EN Words: 0.55M DE Words: 0.54M	Pairs: 91,630 EN Words: 1.13 M ES Words: 1.16 M
0.8	Pairs: 112,341 EN Words: 2.36M RO Words: 2.4 M	Pairs: 119,480 EN Words: 2.08M DE Words: 2.01M	Pairs: 576,179 EN Words: 10.5M ES Words: 11.29M
0.7	Pairs: 142,512 EN Words: 2.99M RO Words: 3.04M	Pairs: 190,135 EN Words: 3.49M DE Words: 3.37M	Pairs: 1,219,866 EN Words: 23.73M ES Words: 25.93M
0.6	Pairs: 169,662 EN Words: 3.58M RO Words: 3.63M	Pairs: 255,128 EN Words: 4.89M DE Words: 4.7M	Pairs: 1,579,692 EN Words: 31.02M ES Words: 33.71M
0.5	Pairs: 201,263 EN Words: 4.26M RO Words: 4.33M	Pairs: 322,011 EN Words: 6.45M DE Words: 6.19M	Pairs: 1,838,794 EN Words: 36.51M ES Words: 39.55M
0.4	Pairs: 252,203 EN Words: 5.42M RO Words: 5.48M	Pairs: 412,608 EN Words: 8.47M DE Words:8.13 M	Pairs: 2,102,025 EN Words: 42.32M ES Words: 45.57M
0.3	Pairs: 317,238 EN Words: 6.89M RO Words: 6.96M	Pairs: 559,235 EN Words: 11.8M DE Words: 11.4M	Pairs: 2,656,915 EN Words: 54.93M ES Words: 58.52M

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

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.

[8] http://dumps.wikimedia.org/backup-index.html of December 22nd, 2012

[9] We deliberately avoid using the term "parallel sentences", because, as will be shown, in our experiments we consider not only parallel sentences.

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 equal or higher than 0.7 are parallel. Those pairs with a translation similarity measure between 0.3 and 0.6 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 dropped them from SMT experiments.

3.2 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 of the evaluation of the SMT systems was only to indicate the best threshold for selecting the training set from the Wiki-Train for subsequent building of the Wiki-Translators. As the standard SMT system we chose Moses surface to surface translation, lexical reordering wbe-msd-bidirectional-fe model, a maximum 4 words phrase-length, and the default values for the rest of parameters.

The target **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 parallel sentence pairs as development set (**dev-set**). These 11,000 sentences were removed from the Wiki-Base (L1, L2) that were meant as training corpora for each language pair. When sampling parallel sentence pairs (test- and dev-sets), we were careful to observe 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 seven training corpora were filtered with Moses' cleaning script with the same restrictions mentioned above. For every language, both the training corpora and the test-set were tokenized using Moses' tokenizer script and true-cased. The quality of the translation systems was 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 test and development set sentences from the training corpora does not ensure an unbiased 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 could be used for identification and elimination from the training set of all the pairs coming from the same documents from which the development and evaluation sets were selected. Yet, due to the nature of the Wikipedia article authoring, even this strategy of filtering the development and evaluation would not ensure an unbiased evaluation. The Wikipedia contributors are given specific instructions for authoring documents [10] and because of complying with these instructions, inevitably 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 all similar (pattern-based) sentences from train-set, 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.

TM based on Wiki-Base	BLEU SCORE RO->EN	BLEU SCORE DE->EN	BLEU SCORE ES->EN
$TM_{[0.3, 1]}$	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

Table 3: Comparison among SMT systems trained on various parts of Wiki-Base

Table 3 summarizes the results of this first step experiment, with the best BLEU scores (bold figures) identifying the most MT useful parts of Wiki-Base (L1,L2). We considered $TM_{[0.7, 1]}$ as the baseline for all language pairs.

3.3 Phase 2: Building Wiki-Train (L1, L2)

The experiments on Wiki-base revealed that the most useful training data was 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-ran GIZA++ on these subsets of Wiki-Base to extract new lexicons.

The new lexicons were merged with the initial ones and the LEXACC extraction was repeated with the resulted mined comparable sentence-pairs denoted as Wiki-Train.

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

Sim. score	EN-RO	EN-DE	EN-ES
0.9	Pairs 66,777 EN Words 1.08M RO Words 1.09M	Pairs 97,930 EN Words 1.07M DE Words 1.04M	Pairs 113,946 EN Words 1.16M ES Words 1.19M
0.8	Pairs 152,015 EN Words 2.69M RO Words 2.7 M	Pairs 272,358 EN Words 3.7M DE Words 3.6M	Pairs 597,992 EN Words 9.73M ES Words 10.51M
0.7	Pairs 189,875 EN Words 3.36M RO Words 3.37M	Pairs 434,019 EN Words 6.2M DE Words 5,93M	Pairs 1,122,379 EN Words 19.94M ES Words 21.82M
0.6	Pairs 221,661 EN Words 3.96M RO Words 3.97M	Pairs 611,868 EN Words 8.94M DE Words 8.53M	Pairs 1,393,444 EN Words 25.07M ES Words 27.41M
0.5	Pairs 260,287 EN Words 4.72M RO Words 4.72M	Pairs 814,041 EN Words 12.36M DE Words 11.79M	Pairs 1,587,276 EN Words 28.99M ES Words 31.57M
0.4	Pairs 335,615 EN Words 6.33M RO Words 6.32 M	Pairs 1,136,734 EN Words 18.09M DE Words 17.31M	Pairs 1,807,892 EN Words 33.62M ES Words 36.37M
0.3	Pairs 444,102 EN Words 8.71 M RO Words 8.70M	Pairs 1,848,651 EN Words 31.41M DE Words 30.18M	Pairs 2,288,163 EN Words 44.02M ES Words 47.18M

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

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 (and eliminated) 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 contains the highest percentage of fully parallel sentence pairs.

3.4 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. 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 point lower, suggesting that removing duplicates was not the best filtering option.

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 of (for in-genre or in-domain experiments).

TM based on Wiki-Train	TM [0.5, 1] RO->EN	TM [0.5, 1] DE->EN	TM [0.3, 1] ES->EN
BLEU SCORE	**41.09**	**40.82**	**46.28**
	TM [0.5, 1] EN -> RO	**TM [0.5, 1] EN -> DE**	**TM [0.3, 1] EN -> ES**
BLEU SCORE	**29.61**	**35.18**	**46.00**

Table 5: Best translation SMT systems, trained on Wiki-Train[11]

4. Comparison with other works

Translation for Romanian-English language pair was also studied in (Boroş et al., 2013; Dumitrescu et al., 2012; 2013) among others. In these experiments we had explicit interests in experiments on using in-domain vs. 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 Romanian->English, similar to those reported here, were surface based 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[12] was translated into English using combinations of:

- a WIKI5-based translation model (240K sentence pairs)/WIKI5-based language model;
- a global translation model (1.7M sentence pairs)/global language model named "ALL", trained on the concatenation of all seven specific corpora.

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

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

Table 6: BLEU scores (RO->EN) on 1000 sentences Wikipedia test-set of Boroş et al. (2013)

Boroş et al. (2013)'s results confirm the conclusion we claimed earlier: the ALL system does not perform better than the in-genre 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-genre for the test data and much larger: the language model was built from entire Romanian Wikipedia (more than 500,000 sentences), 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

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

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

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. Boroş et al. (2013) report BLEU scores on JRC-Acquis test-sets (with domain specific training) almost double than those obtained on Wikipedia test-sets.

The most similar experiments to ours were 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 all the extracted sentence pairs for training, 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 data 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 retrained the two systems on the slightly modified training corpora. Since they used MERT-optimized translation systems in their experiments, we also optimized, by MERT, our new $TM_{[0.5, 1]}$ for German->English and new $TM_{[0.4, 1]}$[13] for Spanish->English translation systems, using the respective dev-sets (each containing 1,000 sentence pairs, as described in Section 3.2).

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. (2010) 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.

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

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

For Spanish-English test-set of Smith et al. (2010) our result is significantly better (6.89 BLEU points) than theirs, in spite of almost 4 times less training data. For the German-English pair, the BLEU score difference between $TM_{[0.5, 1]}$ and Large+Wiki systems is smaller, but still statistically significant (1.34 points), and one should also notice that our system used 10 times less training data.

Surprisingly, 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[14] 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 insight was that the test-sets contained many pieces of texts that looked like section titles. 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 and titles included only 45% passed the LEXACC filtering. As for parallelism status of the sentence pairs in the 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 were a little bit higher (80.8% for Spanish-English and 32.82% for German-English). These evaluations also outline 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[15], while in the German-English the same happens for similarity scores lower than 0.3[16]. The most plausible explanation was that one of the LEXACC's parameters (cross-linking

[13] Although in our earlier experiments on Spanish->English the model $TM_{[0.3, 1]}$ was the best performing on our test data, on the Microsoft test-set the model that achieved the best BLEU score was $TM_{[0.4, 1]}$, exceeding with 0.32 BLEU points the performance of the $TM_{[0.3, 1]}$ model.

[14] 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.

[15] Similarity score (ES-EN) 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.

[16] Similarity score (DE-EN) 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 .

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).

5. 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 the inherent noise they contain might be detrimental to the translation quality.

On the contrary, our experiments demonstrated that in-genre comparable data are strongly preferable to out-of-genre 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-genre translation dictionary is not necessary, and it can be constructed subsequently, starting with a dictionary extracted from whatever parallel 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 which did not include in both parts the same numeric entities (e.g. numbers, dates, and times) was dropped for 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 be likely to 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 (Boroș et al, 2013) and cascading translators (Tufiș & 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), and may significantly reduce the number of out-of-vocabulary input tokens.

The parallel Wiki corpora, including the test-sets (containing 10,000) and the dev–sets (containing 1,000 sentences) are freely available on-line[17].

The archives contain all the extracted sentence pairs, beginning with the similarity threshold 0.1 (see Table 8), thus much more, but, to a large extent, noisy data. Yet, as said before, several useful sentence pairs for each language pair might be recovered.

Sim. score	EN-RO	EN-DE	EN-ES
0.1	Pairs 2,418,227 En Words 63.99M Ro Words 63.96M	Pairs 11,694,784 En Words 295.22M De Words 292.56M	Pairs 3,280,305 En Words 84.30M Es Words 87.37M

Table 8: Full unfiltered data-sets

The LEXACC text miner plus other miners are freely available on ACCURAT project site[18] with the newest version on the RACAI's META-SHARE clone[19].

References

Adafre, S.F. & de Rijke, M. (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.

Boroș, T., Dumitrescu, Ș.D., Ion, R., Ștefănescu, D. & Tufiș, D. (2013). Romanian-English Statistical Translation at RACAI. În E. Mitocariu, M. A. Moruz, D. Cristea, D. Tufiș, M. Clim (eds.) *Proceedings of the 9th International Conference "Linguistic Resources and Tools for Processing the Romanian Language", 16-17 mai,* 2013**,**, Miclăușeni, Romania, 2013. „Alexandru Ioan Cuza" University Publishing House., ISSN 1843-911X, 2013, pp. 81-98.

Dumitrescu, Ș.D., Ion, R., Ștefănescu, D., Boroș, T. & Tufiș. D. (2013). Experiments on Language and Translation Models Adaptation for Statistical Machine Translation. In Dan Tufiș, Vasile Rus, Corina Forăscu (eds.) *TowardsMultilingual Europe 2020: A Romanian Perspective, 2013,* pp. 205-224.

Dumitrescu, Ș.D., Ion, R., Ștefănescu, D., Boroș, T. & Tufiș. D. (2012_. 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.

Gao, Q. & Vogel, S. (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.

Koehn, K. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation, In *Proceedings of the tenth Machine Translation Summit, Phuket, Thailand,* pp. 79-86.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A.

[17] http://dev.racai.ro/dw

[18] http://www.accurat-project.eu/

[19] http://ws.racai.ro:9191/repository/search/?q=lexacc

& Herbst, E. (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.

Mohammadi M., & GhasemAghaee, N. (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,. IEEE Computer Society Washington, DC, USA, pp. 264-268.

Munteanu, D. & Marcu, D. (2005). Improving machine translation performance by exploiting comparable corpora. *Computational Linguistics*, 31(4), pp. 477–504.

Otero, P.G. & Lopez, I.G. (2010). Wikipedia as Multilingual Source of Comparable Corpora. Proceedings of BUCC, Malta, pp. 21-25

Papineni, K., Roukos, S., Ward T., & Zhu, W.J. (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.

Resnik, P. & Smith, N. (2003). The Web as a Parallel Corpus. In Computational Linguistics, vol. 29, no. 3, MIT Press Cambridge, MA, USA, pp. 349-380

Skadiņa, I., Aker, A., Glaros, N., Su, F., Tufis, D., Verlic, M., Vasiljevs, A., & Babych, B. (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

Smith, J.R., Quirk C. & Toutanova, K. (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, © Association for Computational Linguistics (2010), pp. 403-411.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. & Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th LREC Conference*, Genoa, Italy, 22-28 May, 2006, ISBN 2-9517408-2-4, EAN 978-2-9517408-2-2

Ştefănescu, D., Ion R. & Hunsicker, S. (2012). Hybrid parallel sentence mining from comparable corpora. In Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT 2012), Trento, Italy, May 28-30, 2012, pp. 137-144.

Ştefănescu, D., & Ion, R. (2013). Parallel-Wiki: A collection of parallel sentences extracted from Wikipedia. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics*, March 24-30, 2013, Samos, Greece.

Tiedemann, J. (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.

Tufiş, D., Ion, R., Dumitrescu, Ş.D. & Ştefănescu, D. (2013a). 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.

Tufiş, D., Ion, R. & Dumitrescu, Ş.D. (2013b). Wiki-Translator: Multilingual Experiments for In-Domain Translations. In Computer Science Journal of Moldova, vol.21, no.3(63), 2013, pp.1-28.

Tufiş, D., Barbu Mititelu, V., Ştefănescu, D. & Ion, R. (2013c). The Romanian Wordnet in a Nutshell. Language and Evaluation, Springer, Vol. 47, Issue 4, ISSN 1574-020X, DOI: 10.1007/s10579-013-9230-7, pp. 1305-1314.