Wikipedia as an SMT Training Corpus

Dan Tufiș
Institute for AI
Romanian Academy
Bucharest, Romania
tufis@racai.ro

Radu Ion
Institute for AI
Romanian Academy
Bucharest, Romania
radu@racai.ro

Ștefan Daniel Dumitrescu
Institute for AI
Romanian Academy
Bucharest, Romania
sdumitrescu@racai.ro

Dan Ștefănescu
University of Memphis
Memphis, USA
dstfnscu@memphis.edu

Abstract

This article reports on mass experiments supporting the idea that data extracted from strongly comparable corpora may successfully be used to build statistical machine translation systems of reasonable translation quality for in-domain new texts. The experiments were performed for three language pairs: Spanish-English, German-English and Romanian-English, based on large bilingual corpora of similar sentence pairs extracted from the entire dumps of Wikipedia as of June 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.

1 Introduction

Wikipedia is one of the most accessed websites of the Internet according to Alexa.com with a global rank of 6 (being outrun only by major search engines such as Google, Yahoo and Baidu and by Face-book and YouTube). Approximately 14% of all Internet users use it on a daily basis and out of these, more than 50% browse through the English version of Wikipedia which is the most comprehensive one, judged by the number of articles. Wikipedia is not a real parallel corpus, although many documents in different languages are translations from English. Many documents in one language are shortened or adapted translations[1] of 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. In the following, we use the term *MT useful data* to denote sets of bilingual sentences/phrases with a high level of cross-lingual similarity, out of which a word/phrase aligner can extract translation lexicons relevant for the SMT task. SMT engines like Moses (Koehn et al., 2007) produce better translations when presented with larger and larger training parallel corpora. For a given training corpus, it is also known that Moses produces better translations when presented with in-domain new texts (texts from the same domain as the training data, e.g. news, laws, medicine, etc.). 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 their sizes are similar with respect to the domain.

In a previous paper (Ștefănescu et al., 2012) we described in details an open-source parallel data extractor from comparable corpora, developed within the ACCURAT EU-project[4]. Essentially, this extractor allows for identifying similar (translation-wise) sentences in a bilingual comparable corpus. A multi-variable function scores the similarity of each candidate pair, and depending on the level of similarity score (ranging between 0 and 1), one could compile different MT useful data sets. We showed elsewhere (Ion et al., 2011) that with the similarity threshold above 0.7, for all the languages we experimented with, our extracted data, human validated, is really parallel. However, depending on the comparability level of the extraction corpus, the quantity of parallel data extracted may range from 0.1% (weakly comparable corpora) to 29% (strongly comparable corpora) of the entire corpus (Ion et al., 2011). Setting a high similarity threshold has the disadvantage that a significant part of the MT

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

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

[3] JRC-Acquis/DGT Translation Memories are other examples of large parallel juridical texts.

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

Proceedings of Recent Advances in Natural Language Processing, pages 702–709,
Hissar, Bulgaria, 7-13 September 2013.

useful data contained in the comparable corpora is lost.

The experiments we report in this article had multiple purposes:

a) to assess the usefulness of extracted data for SMT by investigating the contribution of less than parallel extracted data to the quality of the translations produced by a baseline SMT; this investigation was driven by iteratively lowering the similarity threshold for the extracted data and evaluating the translation quality for the system trained on the resulted MT useful data.
b) to assess the feasibility of better translating English documents absent from a foreign Wikipedia version; currently, Wikipedia does not offer an integrated translation engine to assist the translation task but this could be a worthy option to consider. With respect to this aim, all our experiments were conducted on in-domain (but unseen during the training) test sets.
c) to add a new domain (for many language pairs) – the encyclopedic domain – to the list of already existing domains for which MT useful data exists (e.g. Tiedemann's OPUS collection multilingual corpora).

In the rest of this paper, after reviewing the related research (Section 2), we provide some statistics on three large sets of similar sentence-pairs extracted from Wikipedia for the English-Spanish, English-German and English-Romanian language pairs (Section 3). In Section 4 we describe the Moses-based experiments with the extracted MT useful data and compare the results with those obtained in a similar scale experiment on Wikipedia. Section 5 describes the follow-up of the previously described experiments with even better results. We conclude with Section 6.

2 Related work

Due to its structure with linked articles on the same subject and because, frequently, articles in foreign languages contain adapted versions of the translations (or just the translation) of the English or other languages counterparts, Wikipedia is arguably the largest strongly comparable corpus available online. It has been the test bed of many attempts at parallel sentence mining.

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 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 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 of 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. Based on binary Maximum Entropy classifiers, in the spirit of Munteanu and 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 distinguish between parallel sentences and non-parallel sentences using 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. 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. Since the work of Smith et al. (2010) is the only one we know of that extracted parallel corpora of similar sizes to ours, we will reserve a detailed comparison with their work in the evaluation section (Section 4.4). Furthermore, they released their English-Spanish and English-German Wikipedia test sets and so, a direct comparison is made possible. Unfortunately, the large amounts of extracted parallel corpora are not available online for the SMT research community.

3 The Extracted Wiki Datasets

Using LEXACC (Ştefănescu et al., 2012) we mined (Ştefănescu and Ion, 2013) for parallel sentence pairs from selected documents belonging to full dumps of English, Romanian, Spanish and German Wikipedias as of December 2012. Table 1 lists, for different similarity scores (**Sim**) 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 (EnW, DeW, RoW, EsW) in the respective sets of sentence pairs. Data extracted with a given similarity score threshold is a proper sub-set of any data extracted with a lower similarity score threshold.

Sim	En-De	En-Ro	En-Es
0.9	P: 38,390	P: 42,201	P: 91,630
	EnW: 0.695 M	EnW: 0.814 M	EnW: 1.126 M
	DeW: 0.543 M	RoW: 0.828 M	EsW: 1.158 M
0.8	P: 119,480	P: 112,341	P: 576,179
	EnW: 2.077 M	EnW: 2.356 M	EnW: 10.504 M
	DeW: 2.010 M	RoW: 2.399 M	EsW: 11.285 M
0.7	P: 190,135	P: 142,512	P: 1,219,866
	EnW: 3.494 M	EnW: 2.987 M	EnW: 23.730 M
	DeW: 3.371 M	RoW: 3.036 M	EsW: 25.931 M
0.6	P: 255,128	P: 169,662	P: 1,579,692
	EnW: 4.891 M	EnW: 3.577 M	EnW: 31.022 M
	DeW: 4.698 M	RoW: 3.634 M	EsW: 33.706 M
0.5	P: 322,011	P: 201,263	P: 1,838,794
	EnW: 6.453 M	EnW: 4.262 M	EnW: 36.512 M
	DeW: 6.186 M	RoW: 4.325 M	EsW: 39.545 M
0.4	P: 412,608	P: 252,203	P: 2,102,025
	EnW: 8.470 M	EnW: 5.415 M	EnW: 42.316 M
	DeW:8.132 M	RoW: 5.482 M	EsW: 45.565 M
0.3	P: 559,235	P: 317,238	P: 2,656,915
	EnW: 13.740M	EnW: 6.886 M	EnW: 54.932 M
	DeW: 11.353M	RoW: 6.963 M	EsW: 58.524 M
0.2	P: 929,956	P: 449,640	P: 3,850,782
	EnW: 25.485M	EnW: 9.956 M	EnW: 88.567 M
	DeW: 21.492M	RoW:10.056 M	EsW: 93.047 M
0.1	P: 1,279,166	P: 683,223	P: 5,025,786
	EnW: 37.076M	EnW: 16.275 M	EnW: 122.760 M
	DeW: 31.537M	RoW:.16.420 M	EsW: 128.132 M

Table 1: Number of parallel sentences and words extracted for each language pair, for a given threshold (Ştefănescu and Ion, 2013)

From Table 1, one could easily calculate the average word length for the extracted sentences for each language and each threshold value. It is not surprising that longer the sentences their similarity scores get lower. For the En-De language pair, the sentence word length varied for En from 28.98 to 18.11 while for De it varied from 24.65 to 14.43. A similar variation may be noticed for En-Es pair: from 24.42 to 12.28 (En) and from 25.49 to 12.63 (Es). For En-Ro the average sentence word length varied less: from 23.82 to 19.27 (En) and from 24.03 to 19.63 (Ro).

By random manual inspection of the generated sentence pairs, we confirmed earlier evaluations (Ion et al., 2011) that, in general, irrespective of the language pair, sentence pairs with a translation similarity measure of at least 0.7 are entirely parallel (e.g. "In 2003, Africa 2 Africa was merged with SABC Africa." ⇔ "En 2003, Africa 2 Africa fue fusionada con SABC Africa.", score 0.97), those with a translation similarity measure of at least 0.5 have extended parallel fragments which an accurate word or phrase aligner easily detects (e.g. "Besides regular repairs of the existing runways, Prague Airport (Letiště Praha s.p." ⇔ "Además de las habituales refacciones de las pistas, Letiště Praha s.p.", score 0.59). Below 0.5, sentences usually become strongly comparable. 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 (e.g. "Slaves were previously introduced by the British and French who colonized the island in the 18th century."⇔ "Los esclavos ya habían sido introducidos un siglo antes por los británicos y franceses que trataron de conquistar la isla.", score 0.29). The noisiest data sets were extracted for the 0.1 similarity threshold and we drop them from further experiments.

4 SMT experiments with Wiki datasets

There is a strong opinion, empirically supported, that parallel data extracted from comparable corpora leads to improvements of the translation quality of a baseline MT system when it incorporates this data. This has been exemplified by showing that a baseline MT system trained on data covering one or more domains, when tested on texts out of the respective domain(s), performed significantly worse. Translation models adaptation with data extracted from comparable corpora from the test domain improved the translation quality, but in general not reaching the same quality as in the baseline MT translation of the in-domain texts. One can naturally raise the following question: given a large and continuously growing multilingual collection of documents (such as Wikipedia) what would be a good approach for enhancing a SMT trained to translate Wikipedia-like documents (let's call it Wiki-translator)? The question calls to the limited

available in-domain parallel data for any language pair (the sizes of pair-wise parallel Wikipedias are limited, even for the best represented languages) but suggest the benefits of in-domain adaptation by using comparable data extracted from Wikipedia. This issue is placed into operational terms, by asking the question: what level of sentences comparability is useful for improving the quality of Wiki-translator's output? The experiments described in this section try to provide some hints to the questions above.

We argued that with a high value (0.7) for the similarity threshold, the extracted sentence-pairs can safely be considered truly parallel. However, in Table 1, we showed that the number of sentences pairs with a similarity score of at least 0.7 represents a small portion (ranging from 14% to a maximum of 24%) of the potentially MT useful sentence pairs (corresponding to the threshold 0.1) from the interlinked documents.

In what follows, we give experimental insights by observing how translation improves/degrades when training on parallel sentences with different translation similarity thresholds.

4.1 Experimental setup

As mentioned in Section 4, the English, German and Spanish Wikipedias are the largest ones with substantial cross-lingual coverage. Romanian Wikipedia is medium-sized but containing many translations or adaptations of articles from other languages (mainly English). Consequently, we could find in En-De, En-Es and En-Ro Wikipedias a number of parallel sentences (190,135 for En-De with more than 6.86 million words, 142,512 for En-Ro with more than 6 million words and 1,219,866 for En-Es with almost 50 million words) allowing for building baseline Wiki-translators for these language pairs. The large sets of comparable sentences allowed us to conduct experiments on assessing the translation quality improvement/degradation when the parallel core training corpora were gradually extended with comparable but less and less parallel sentence pairs.

As the standard SMT system we chose Moses[5] with the default parameters for factorial optimization. We used it with the following parameters:

- surface-to-surface translation;
- phrase length of maximum 4 words;
- lexical reordering model with parameters `wbe-msd-bidirectional-fe`.

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

The **language model** (LM) for all experiments was trained on entire monolingual, sentence-split English Wikipedia, after removing the administrative articles as described in Section 4. The language model was limited to 5-grams and the counts were smoothed with the interpolated Knesser-Ney method.

The **test sets for the three language pairs** were created by concatenating randomly extracted 2500 sentence pairs from each similarity interval ensuring parallelism ([0.6, 1], [0.7, 1], [0.8, 1] and [0.9, 1]). The sentence pairs extracted from each similarity interval were manually checked for parallelism. Thus we obtained 10,000 parallel sentence pairs for each language pair. These sentences were removed from the training data. In compiling the test 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.

Once the test sets were ready, we further trained **eight translation models** (TM), for each language pair, over cumulative threshold intervals beginning with 0.2: $TM_{[0.2,\ 1]}$ for [0.2, 1], $TM_{[0.3,\ 1]}$ for [0.3, 1] …, $TM_{[0.9,\ 1]}$ for [0.9, 1]. The training data for $TM_{[0.2,\ 1]}$ was the largest but the noisiest, while the training data for $TM_{[0.9,\ 1]}$ was the smallest but fully parallel. 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.

We are interested in finding out if the quality of the translation system based on the translation model TM_i were significantly different from the quality of the translation system based on the translation model TM_{i+1}, where TM_i and TM_{i+1} are translation models built as described in the previous sub-section. The quality of the translation systems was measured as usual in terms of their BLEU score (Papineni et al., 2002) on the same test data (10,000 parallel sentence pairs).

4.2 SMT results for Spanish-English and German-English

Table 2 shows the variations of the BLEU scores on the Spanish-English test set for the SMTs with different translation models. The shaded lines indicate the translation models built on fully parallel data. The better score of $TM_{[0.7,\ 1]}$ as compared to those of $TM_{[0.8,\ 1]}$ and $TM_{[0.9,\ 1]}$ is not surprising: the parallel training data is signifi-

cantly larger: 190,135 pairs for $TM_{[0.7, 1]}$, 119,480 pairs for $TM_{[0.8, 1]}$ and only 38,390 for $TM_{[0.9, 1]}$. However, with additionally more 369,100 sentence pairs less parallel, $TM_{[0.3, 1]}$ achieves the best performance, with an statistically significant increase of 0.31 BLEU points and a much larger lexical coverage.

One can further see from Table 2, that in spite of the major reduction of the size of the training data, a significant increase in the BLEU score is achieved from the 0.2 translation model to 0.3.

The explanation is that most of the eliminated data was noisy; the training corpus became cleaner. This is a clear indication that comparable data existing in the respective training sets: **1)** does not degrade SMT performance; **2)** it makes the translation model more robust.

TM	BLEU SCORE
$TM_{[0.2, 1]}$	47.22
$TM_{[0.3, 1]}$	**47.59**
$TM_{[0.4, 1]}$	47.52
$TM_{[0.5, 1]}$	47.53
$TM_{[0.6, 1]}$	47.44
$TM_{[0.7, 1]}$	47.28
$TM_{[0.8, 1]}$	46.27
$TM_{[0.9, 1]}$	39.68

Table 2: Experimental SMT results on Es-En

Similar comments can be made for the English-German experiment. Table 3 presents the experimental results. This time the best BLEU score is obtained using $TM_{[0.5, 1]}$.

TM	BLEU SCORE
$TM_{[0.2, 1]}$	37.61
$TM_{[0.3, 1]}$	39.16
$TM_{[0.4, 1]}$	39.46
$TM_{[0.5, 1]}$	**39.52**
$TM_{[0.6, 1]}$	39.5
$TM_{[0.7, 1]}$	39.24
$TM_{[0.8, 1]}$	38.57
$TM_{[0.9, 1]}$	34.73

Table 3: Experimental SMT results on De-En

4.3 SMT results for Romanian-English

Translation for Romanian-English language pair has also been studied in Dumitrescu et al. (2013) with explicit interest for the in-domain/out-of-domain test/train data, using Moses in various configurations for surface-to-surface and factored translation. Out of the seven domain specific corpora (legal, transcribed speech, parliamentary debates, literature, medicine, news and encyclopedic) the encyclopedic corpus was based on Wikipedia. They have experimented with English-Romanian parallel sentence pairs extracted from Wikipedia using LEXACC at a fixed threshold: 0.5 (called "WIKI5"). A random selection of unseen 1000 Wikipedia Romanian test sentences has been 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", made by concatenating all specific corpora.

Table 4 gives the details, giving the BLEU scores for the Moses configuration similar to ours: surface-to-surface translation, with the language/translation model combinations described above.

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

Table 4: BLEU scores on 1000 sentences Wikipedia test set of Dumitrescu et al. (2013)

Dumitrescu et al.'s results confirm the conclusion we claimed earlier: the ALL system performs worse than the in-domain WIKI5 system.

Our present results show the same characteristics as those of the Spanish-English and German-English experiments presented earlier. They are summarized in Table 5.

TM	BLEU SCORE
$TM_{[0.2, 1]}$	36.1
$TM_{[0.3, 1]}$	37.24
$TM_{[0.4, 1]}$	37.71
$TM_{[0.5, 1]}$	**37.99**
$TM_{[0.6, 1]}$	37.85
$TM_{[0.7, 1]}$	37.39
$TM_{[0.8, 1]}$	36.89
$TM_{[0.9, 1]}$	32.76

Table 5: Experimental SMT results on Ro-En

The almost eight BLEU points difference between our results and those in (Dumitrescu et al., 2013) may be explained by:

1) our language model was entirely in-domain for the test data and much larger: our language model was built from entire Romanian Wikipedia (more than 220,000 documents) while the language model in (Dumitrescu et al., 2013) was built only from the Romanian doc-

ument paired to English documents (less than 100,000 documents);

2) different Moses filtering parameters (e.g. the length filtering parameters),

3) different test sets.

4.4 Comparison with Smith et al. (2010)

As mentioned in Section 2, Smith et al. (2010) mined for parallel sentences from Wikipedia producing parallel corpora of sizes similar to ours Furthermore, they have made their Wikipedia test set available for Spanish-English and German-English (500 sentence pairs per language pair). We have translated these test sets (after being true-cased) with our best translation models (0.3 for Spanish-English and 0.5 for German-English) and also with Google Translate (as of mid-February 2012). Table 6 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. "0.3 TM" and "0.5 TM" are our 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	System	BLEU
Spanish-English	9642K	Large+Wiki	43.30
	2288K	TM $_{[0.4, 1]}$	**50.19**
	N/A	Google	44.43
German-English	8388K	Large+Wiki	23.30
	306K	TM $_{[0.5, 1]}$	**23.34**
	N/A	Google	21.64

Table 6: Comparison between SMT systems on the Wikipedia test set provided by Smith et al. (2010)

It is thus empirically supported the finding that indiscriminately adding more out-of domain data, when large enough in-domain data already exists (as in these compared experiments), produces worse results.

5 Bootstrapping experiments

The astute reader may have noticed that the dictionaries used by LEXACC for mining MT useful data were extracted by GIZA++ from out-of-domain corpora (JRC-Acquis and Europarl). After obtaining the sets of in-domain MT useful data for the three language pairs discussed above, it was a natural decision to go one step further: compute new translation dictionaries by merging the old ones with the dictionaries generated by GIZA++ from in-domain data (extracted as described in Section 4) and re-do the SMT experiments described in Section 5. Since the full chain of experiments for the three language pairs is extremely time consuming, at the time of this writing we have the new results only for En-Ro language pair, which has the smallest datasets.

5.1 English-Romanian new extracted data

The earlier experiments empirically showed that the Similarity Score below 0.2 produced too much noisy data to be useful in SMT experiments. Therefore, we proceed with the LEXACC extraction process considering Similarity Score (**Sim**) higher or equal to 0.2.

Table 7 shows a significant increase of the number of extracted bilingual sentence pairs when the out-of-domain translation dictionary is extended by the in-domain translation lexicon.

Sim	Initial En-Ro	Boosted En-Ro
0.9	P: 42,201 EnW: 0.814 M RoW: 0.828 M	P: 66,777 EnW: 1.077 M RoW: 1.085 M
0.8	P: 112,341 EnW: 2.356 M RoW: 2.399 M	P: 152,015 EnW: 2.688 M RoW: 2.698 M
0.7	P: 142,512 EnW: 2.987 M RoW: 3.036 M	P: 189,875 EnW: 3.364 M RoW: 3.372 M
0.6	P: 169,662 EnW: 3.577 M RoW: 3.634 M	P: 221,661 EnW: 3.961 M RoW: 3.970 M
0.5	P: 201,263 EnW: 4.262 M RoW: 4.325 M	P: 260,287 EnW: 4,715 M RoW: 4,722 M
0.4	P: 252,203 EnW: 5.415 M RoW: 5.482 M	P: 335,615 EnW: 6.329 M RoW: 6.324 M
0.3	P: 317,238 EnW: 6.886 M RoW: 6.963 M	P: 444,102 EnW: 8.712 M RoW: 8.700 M
0.2	P: 449,640 EnW: 9.956 M RoW:10.056 M	P: 811,113 EnW: 171.425 M RoW: 171.109 M

Table 7: Boosting: comparison between the number of parallel sentences and words extracted for En-Ro

The new extracted corpus was used for the similar SMT experiments as described in Section 5. The test set was selected from completely parallel documents, not contained into the data extraction space. We changed the test set construction strategy using entire parallel documents and not sentence pairs from the parallel documents.

The first strategy could be suspected of biasing, since the contexts of the tested sentences (the documents from where the test sentence-pairs were extracted) were used for training.

The test set contains 1,000 Ro-En parallel sentences. Table 8 shows the results.

Again, we outline the differences in BLEU scores for the initial SMT experiments and the boosted ones.

TM	Initial BLEU score	Boosted BLEU score
TM $_{[0.2, 1]}$	36.10	47.31
TM $_{[0.3, 1]}$	37.24	49.83
TM $_{[0.4, 1]}$	37.71	49.83
TM $_{[0.5, 1]}$	**37.99**	50.74
TM $_{[0.6, 1]}$	37.85	**50.78**
TM $_{[0.7, 1]}$	37.39	50.52
TM $_{[0.8, 1]}$	36.89	49.85
TM $_{[0.9, 1]}$	32.76	45.52

Table 8: Boosting: BLEU comparisons on Ro-En

We made also translation experiments for the other direction, Ro-En, and as expected the translation accuracy (in terms of BLEU scores) was significantly lower. The best BLEU score for En-Ro translation direction was **44.09**, but this time for the translation model trained on the bilingual corpus with the similarity score equal or higher than 0.5 (TM $_{[0.6, 1]}$).

The last step in our experimental chain was to optimize the translation parameters using the usual MERT procedure. The development set used to tune the translation parameters had 1,000 parallel sentences, not used in the training or test sets. Not surprisingly, the BLEU scores further improve. Table 9 summarizes the new results:

TM	Boosted BLEU score	MERT Boosted BLEU score
TM $_{[0.2, 1]}$	47.31	48.92
TM $_{[0.3, 1]}$	49.83	50.61
TM $_{[0.4, 1]}$	49.83	50.48
TM $_{[0.5, 1]}$	50.74	**51.05**
TM $_{[0.6, 1]}$	**50.78**	50.97
TM $_{[0.7, 1]}$	50.52	50.65
TM $_{[0.8, 1]}$	49.85	50.65
TM $_{[0.9, 1]}$	45.52	46.69

Table 9: Optimized Boosting: BLEU comparisons on Ro-En

So far, we obtained our best result of 51.05 BLEU for the Ro-En direction, using the MERT-enhanced Boosted method.

6 Conclusions

We have shown that Wikipedia is a rich resource for parallel sentence mining in Statistical Machine Translation. Comparing different translation models containing MT useful data ranging from comparable, through 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. The parallel Wiki corpora (before and after the boosting step), including the two test sets (containing 10,000 and respectively 1,000 sentences) are freely available on-line[6]. We want to clarify one aspect though: it is not the case that our extracted data is the maximally MT useful data. We evaluated and extracted only full sentences. A finer-grained (sub-sentential) extractor would likely generate more MT useful data.

Acknowledgments

This work has been supported by the EU under the Grant Agreements no. 248347 and no. 270893.

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

References

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.

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.

Radu Ion, Mārcis Pinnis, Gregor Thurmair, Ahmet Aker, Rob Gaizauskas, Mateja Verlic, and Nikos Glaros. 2011. Extracted data for translation models of SMT and RBMT lexicon from aligned comparable corpora. Deliverable D2.5 of the ACCURAT Project. Available online at http://www.accurat-project.eu/index.php?p=deliverables

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej 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, 177-180.

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

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

Kishore Papineni, Salim Roukos, Todd Ward and 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.

Jason R. Smith, Chris Quirk and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In Proceedings of *the 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).

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th LREC Conference*, Genoa, Italy, 22-28 May, 2006, pp.2142-2147, ISBN 2-9517408-2-4, EAN 9782951740822

Dan Ştefănescu, Radu Ion and 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

Dan Ştefănescu, and Radu Ion. 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 (CICLING 2013)*, March 24-30, 2013, Samos, Greece.

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, pp. 2214—2218.