Parallel-Wiki: A Collection of Parallel Sentences Extracted from Wikipedia

Dan Ştefănescu[1,2] and Radu Ion[1]

[1] Research Institute for Artificial Intelligence, Romanian Academy
{danstef,radu}@racai.ro
[2] Department of Computer Science, The University of Memphis
dstfnscu@memphis.edu

Abstract. Parallel corpora are essential resources for certain Natural Language Processing tasks such as Statistical Machine Translation. However, the existing publically available parallel corpora are specific to limited genres or domains, mostly juridical (e.g. JRC-Acquis) and medical (e.g. EMEA), and there is a lack of such resources for the general domain. This paper addresses this issue and presents a collection of parallel sentences extracted from the entire Wikipedia collection of documents for the following pairs of languages: English-German, English-Romanian and English-Spanish. Our work began with the processing of the publically available Wikipedia static dumps for the three languages involved. The existing text was stripped of the specific mark-up, cleaned of non-textual entries like images or tables and sentence-split. Then, corresponding documents for the above mentioned pairs of languages were identified using the cross-lingual Wikipedia links embedded within the documents themselves. Considering them comparable documents, we further employed a publically available tool named LEXACC, developed during the ACCURAT project, to extract parallel sentences from the preprocessed data. LEXACC assigns a score to each extracted pair, which is a measure of the degree of parallelism between the two sentences in the pair. These scores allow researchers to select only those sentences having a certain degree of parallelism suited for their intended purposes. This resource is publically available at:
http://ws.racai.ro:9191/repository/search/?q=Parallel+Wiki

Keywords: Parallel Data, Comparable Corpora, Statistical Machine Translation, Parallel Sentence Extraction for Comparable Corpora

1 Introduction

During recent years, Statistical Machine Translation (SMT) has received a lot of attention from the scientific community, attracting more and more researchers. Some of this interest is due to companies like Google or Microsoft, whose public SMT engines attract a great deal of curiosity and shape the belief that building an SMT system for informative translations that is widely accepted by Internet users is very near (increasing the level of awareness of the field). However, much of the research in this direction makes use of the same SMT model (Shannon's noisy channel) with its very popular implementation, the Moses SMT Toolkit (Koehn et al., 2007). So far, Moses has been proven to be the best publically available engine on which SMT systems are built. The differences in quality between such systems come to depend on the resources used by the Moses decoder or on the post-processing steps which aim to correct some of its mistakes. Still, in terms of procuring some of the needed resources like translation models, the scientific community has very few publically available options even for resource-rich languages. In order to build translation models, one needs parallel text aligned at the sentence level and such resources cannot be easily acquired in large quantities. Most of the available ones are juridical, medical or technical collections of documents, which are often the result of efforts beyond the NLP field. For example, JRC-Acquis[1] (Steinberger et. al, 2006) is a well-known collection of juridical parallel texts in 22 languages covering the EU legislation. It is the most used parallel corpus for Statistical Machine Translation experiments. OPUS[2] (Tiedemann, 2012) is a collection of parallel corpora that includes many known freely available such resources. Some of them are: (i) EUROPARL (European Parliament Proceedings) (Koehn, 2005), (ii) EUconst (the European constitution), which are both juridical texts, (iii) EMEA (European Medicines Agency documents) which belongs to the medical domain, several technical parallel texts like (iv) ECB (European Central Bank corpus), (v) KDE4 localization files, (vi) KDE manual corpus, (vii) PHP manual corpus, etc., some subtitles corpora like (viii) OpenSubs or (ix) TEP (Tehran English-Persian subtitle corpus) and news corpora like SETIMES (parallel news corpus of the Balkan languages).

From the above enumeration of existing parallel texts, one can infer that the general domain is poorly covered and more than this, there are languages for which parallel texts are scarce, no matter the domain. This is why the research community started to explore the possibility of acquiring parallel data from comparable texts. Such texts contain documents referring to the same subject or topic, but are not reciprocal translations. The problem of extracting data from comparable corpora began to be studied in the late 90s, as soon as people realized that the Web can be seen as a vast source of comparable documents. Among the important contributions to this area of research we have to mention the works of Wu (1994), Zhao and Vogel (2002), Resnik and Smith (2003), Fung and Cheung (2004), Munteanu and Marcu (2005), Quirk et al. (2007) and Tillmann (2009). Recent research includes that of Rauf and Schwenk (2011) and Ştefănescu et al. (2012). The most recent European projects on this topic are ACCURAT[3] (Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation) and TTC[4] (Terminology extraction, Translation tools and Comparable corpora).

Our own experiments on mining parallel data from comparable corpora were conducted within the ACCURAT project and led to the development of a tool named

[1] http://ipsc.jrc.ec.europa.eu/index.php?id=198

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

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

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

LEXACC[5]. With LEXACC, which is thoroughly described in Ştefănescu et al. (2012), one can extract parallel sentences from comparable corpora even if the level of comparability is low.

This paper describes the process of using LEXACC for harvesting parallel sentences from Wikipidia[6], for three language pairs: English-German, English-Romanian and English-Spanish. The next section presents related work, the following one gives information about the Wikipedia data we considered and the pre-processing steps we undertook in order to clean it. Section 3 details the procedure we followed for extracting the data, while section 4 contains statistics about the newly created resources. The paper ends with conclusions and ideas for further research.

2 Related Work

Considering that the largest existing publically available database of comparable documents is Wikipedia, a natural step would be to use it for harvesting parallel data. Adafre and Rijke (2006) are among the first to follow this idea, working on English-Dutch pair of languages. They suggested two approaches. The first one employs an MT system to generate a rough translation of a page and then uses word overlap between sentences as a similarity measure. In the second approach, the similarity measure is likewise computed, but this time the sentences are represented by entries in a shared lexicon built on concepts and entities that have entries in Wikipedia. Adafre and Rijke conducted small-scale experiments on a random sample of 30 Wikipedia page pairs. To find the parallel sentences, they considered the entire Cartesian product of source-target sentence pairs, an approach which is not feasible when dealing with datasets many orders of magnitude larger.

Yasuda and Sumita (2008) proposed a framework of a Machine Translation (MT) bootstrapping method on multilingual Wikipedia articles. According to the authors, this method can "simultaneously generate a statistical machine translation and a sentence-aligned corpus." They conducted their experiments on the Japanese-English language pair, working with a 2007 Wikipedia version. As stated by the authors, at that time, the on-line encyclopedia contained 1,500 sentences for Japanese and 50,000,000 for English and they also considered the entire Cartesian product of these sets in the process of finding similar sentence pairs.

Smith at al. (2010) exploited the observation that parallel sentences pairs are frequently found in close proximity and attempted to model the document level alignment accordingly. To do that, they used four categories of features: (i) features derived from word alignments, (ii) distortion features, (iii) features derived from Wikipedia markup and (iv) word-level induced lexicon features. Smith at al. worked on three language pairs (i.e. Spanish-English, Bulgarian-English and German-English), conducting small-scale experiments on 20 Wikipedia article pairs for each language pair. While their datasets are no longer available at the URL provided within the paper, their work showed that Wikipedia is a useful resource for mining parallel data.

[5] http://nlptools.racai.ro/nlptools/index.php?page=lexacc

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

Mohammadi and Ghasem-Aghaee (2010) improved the method of Adafre and Rijke (2006) by making use of Gale and Church (1993) observation that “longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences.” Consequently, they reduced the search space represented by the Cartesian product between the sets containing the source and target sentences within an article pair. They experimented with different similarity measures between candidate pairs: Dice, Cosine and Jaccard coefficients, the latter obtaining the best results. Mohammadi and Ghasem-Aghaee evaluated their method on 30 Wikipedia article pairs and constructed a Persian-English parallel corpus by mining 1,600 article pairs.

Another work we have to mention is that of Birch et al. (2011) who released the Indic multi-parallel corpus in December, 2011. This corpus contains about 2,000 Wikipedia sentences translated into 6 Indic languages. As mentioned by the authors, “the data was translated by non-expert translators hired over Mechanical Turk and so it is of mixed quality.” To our knowledge, this is the only publically available resource of parallel sentences extracted (though not automatically) from Wikipedia.

3 Wikipedia Data

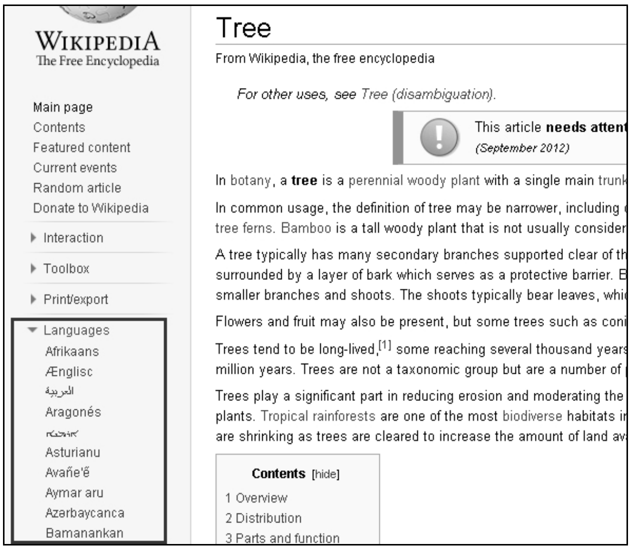
Given the above amount of research dedicated to extracting parallel data from Wikipedia, it is rather unexpected that such publically available resources are virtually non-existent. The main reason for this absence is probably the high amount of computing resources (both in time and memory) necessary to run the proposed algorithms. This is why we have considered leveraging our previous work on this subject (Ştefănescu et al., 2012) and using LEXACC for harvesting parallel data from the entire Wikipedia. Our goal is to provide a large collection of parallel sentences to the research community for three language pairs: English-German, English-Romanian and English-Spanish. Three of the involved languages (i.e. English, German and Spanish) are listed among those having the top number of Wikipedia articles, while Romanian is also richly represented with almost 200,000 articles.

“Wikipedia is a free, collaboratively edited, and multilingual Internet encyclopedia supported by the non-profit Wikimedia Foundation.” (cf. http://en.wikipedia.org/wiki/ Wikipedia). At the moment this paper was written (December 2012), it had over 4.1 million English articles containing embedded links to articles on the same subjects (see Fig. 1), but in different languages. According to Wikipedia, in December 2012 there were 285 languages for which it contained articles, making it the largest publically available collection of comparable documents.

Wikipedia articles can be downloaded by going to the URL[7] containing the so-called “database backup dumps”. Wikipedia states that these dumps contain “a complete copy of all Wikimedia wikis, in the form of wikitext source and metadata embedded in XML”. During May and June 2012, we downloaded the dumps having the label “current versions only” for English (e.g. file enwiki-20120601-pages-meta-

[7] http://dumps.wikimedia.org/backup-index.html

current.xml.bz2), German, Romanian and Spanish (see Table 1 for quantitative details). Parsing the English XML dump, we kept only the “proper” articles containing links to their corresponding articles in the other three languages. By proper articles we mean those that are not *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.



Tree

From Wikipedia, the free encyclopedia

For other uses, see Tree (disambiguation).

This article **needs attent**
(September 2012)

In botany, a **tree** is a perennial woody plant with a single main trunk

In common usage, the definition of tree may be narrower, including
tree ferns. Bamboo is a tall woody plant that is not usually consider

A tree typically has many secondary branches supported clear of th
surrounded by a layer of bark which serves as a protective barrier. E
smaller branches and shoots. The shoots typically bear leaves, whi

Flowers and fruit may also be present, but some trees such as coni

Trees tend to be long-lived,[1] some reaching several thousand years
million years. Trees are not a taxonomic group but are a number of

Trees play a significant part in reducing erosion and moderating the
plants. Tropical rainforests are one of the most biodiverse habitats i
are shrinking as trees are cleared to increase the amount of land av

Contents [hide]
1 Overview
2 Distribution
3 Parts and function

Fig. 1. Most Wikipedia articles contain links (lower-left corner in this image) to articles covering the same subject (e.g. Tree), but written in many different languages

Most of the proper articles do not contain colon characters within their titles, but not all of them (e.g. Blade Runner 3: Replicant Night). Each such article was processed using regular expressions to remove the XML mark-up in order to keep only the raw text, which was saved into a separate file. The non-textual entries like images or tables were also cleaned off. The articles identified as the corresponding articles in the other languages received the same treatment.

Table 1. Figures about the Wikipedia dumps, considering only proper articles

Language	English	German	Romanian	Spanish
Wiki dump date	June 01	May 08	June 10	June 11
Archive size on disk	15.4 Gb	3.9 Gb	277.9 Mb	2.0 Gb
Unpacked XML size	77.6 Gb	15.3 Gb	1.5 Gb	9.6 Gb
# Documents	3,975,895	1,406,603	180,234	894,378

We ended up with the lists of all comparable documents for all our language pairs and the documents themselves, containing only raw text. Since the Romanian Wikipedia had the fewest articles (180,234: 22 times less than the English one and almost 8 times less than the German one), the list of comparable documents for English – Romanian was also the shortest, containing almost 6 times less pairs than the English – German list (see Table 2).

Table 2. The number of comparable documents identified for each considered pair of languages

Language Pair	# Comparable Documents	Size on disk
English-German	715,555	2.8 Gb (English) 2.3 Gb (German)
English-Romanian	122,532	778.1 Mb 198.9 Mb
English-Spanish	573,771	2.5 Gb 1.5 Gb

4 Extracting parallel sentences

With all the comparable documents in place, the next step was to employ LEXACC for harvesting parallel sentences. As input, this tool requires lists of source and target documents and, for a better accuracy, their correspondence. In order to ease its job, we first split the documents into sentences using a freely available sentence splitter[8] based on a Maximum Entropy classifier (Tufiş et al., 2008). This tool uses features that are language independent and though not as accurate as a language-aware sentence splitter, it achieves good results on most Indo-European languages. Furthermore, with the purpose of reducing LEXACC's running time, we partitioned the lists of mapped documents that had to be fed to it into smaller lists containing no more than 50,000 pairs. Consequently, we had 15 such lists for English-German, 3 for English-Romanian and 12 for English-Spanish. LEXACC was run for all these sub-lists in both directions (since its results are not symmetrical) for each language pair. We used as supplementary parameters (i) a flag signaling that the documents were already sentence split and (ii) that we want to keep all sentence pairs for which the assigned translation score was greater than 0.1. The tool was run on an x64 machine having an Intel Core I7 CPU @ 3.33 GHz and 16 GB of RAM. To run LEXACC for an English-German list (list.txt) of document pairs, one needs to use the following command line:

```
lexacc64.exe --docalign list.txt --source en --target de
--output results.txt --param seg=true --param t=0.1
```

LEXACC running time depends on both the size of the collection, but also on the size and quality of the dictionaries it uses as resources. This is why, even if the number of comparable documents was smaller, the running time for English-Romanian

[8] http://nlptools.racai.ro/nlptools/index.php?page=pwiki

was comparable to the others. Table 3 shows the running time needed by LEXACC to extract the parallel sentences. The English-Romanian dictionaries were extracted from the JRC-Acquis corpus and complemented with translation pairs from the Princeton WordNet (Fellbaum, 1998) to Romanian WordNet conceptual alignment (Tufiş et al., 2008). Every English word belonging to a synset is paired with all Romanian words in the corresponding synset and all inflectional variants of the two translation equivalents are also generated.

Table 3. LEXACC running time and the size of dictionaries used as resources

Language Direction	Running time		Dictionary size on disk
	Minutes	Days	
English-German	11,949	8.29	13.6 Mb
German-English	8,973	6.23	13.0 Mb
English-Romanian	8,583	5.96	283.4 Mb
Romanian-English	2,296	1.59	283.4 Mb
English-Spanish	9,786	6.79	15.0 Mb
Spanish-English	8,955	6.21	18.7 Mb
Total	50,542	35.07	657.1 Mb

We used the default LEXACC resources (dictionaries automatically extracted with GIZA++ (Och and Ney, 2003) from the JRC-Acquis). These resources were not available for English-Spanish and we also applied GIZA++ (with the standard parameterization) on both Europarl and JRC-Acquis parallel corpora to obtain them.

For the whole exercise, LEXACC's running time only exceeds one month, but still, this is a short time given the number of comparable documents to be analyzed.

Having parallel sentences in both directions for all language pairs, for each such pair we computed the union of the two sets of data (source-target and target-source), keeping the larger score for those sentence pairs appearing in both sets. This strategy is supported by the facts that LEXACC assigns high translation scores only if certain criteria for determining the translation cohesion are met, and, more often than not, the information needed in order to meet these criteria is not necessarily found in both directions. Moreover, LEXACC's translation similarity measure is tuned to achieve a better precision at the expense of recall and thus, keeping the maximum of the translation similarity score of a sentence pair discovered from both directions ensures the growth of the final parallel dataset. Finally, every sentence pair occurs only once within the entire merged collection (duplicates were eliminated).

All the sentence pairs we extracted are publically available and can be downloaded from http://ws.racai.ro:9191/repository/search/?q=Parallel+Wiki.

5 Statistics of the Extracted Parallel Corpora

The number of parallel sentences extracted by LEXACC is remarkable (see Tables 4 and 5). It is important to notice that the quantity of data acquired for English-Spanish is far greater than the others. This might be explained by the fact that Spanish articles contain much more translations from English documents (or vice-versa). In total, we ended up with 429.7 Mb of data for English-German, 214.5 Mb for English-Romanian and 1.5 Gb for English-Spanish.

Table 4. Number of parallel sentences extracted for all language pairs at different thresholds

Score	English-German	English-Romanian	English-Spanish
0.9	38,390	42,201	91,630
0.8	119,480	112,341	576,179
0.7	190,135	142,512	1,219,866
0.6	255,128	169,662	1,579,692
0.5	322,011	201,263	1,838,794
0.4	412,608	252,203	2,102,025
0.3	559,235	317,238	2,656,915
0.2	929,956	449,640	3,850,782
0.1	1,279,166	683,223	5,025,786

Table 5. The total number of words (alpha-numeric tokens) in the parallel sentences extracted for all language pairs

Score	English-German		English-Romanian		English-Spanish	
	English	German	English	Romanian	English	Spanish
0.9	553,967	543,126	813,595	828,448	1,125,621	1,158,173
0.8	2,076,963	2,010,170	2,355,819	2,399,120	10,503,793	11,285,236
0.7	3,494,316	3,370,622	2,986,957	3,036,061	23,729,717	25,793,126
0.6	4,891,202	4,697,714	3,576,837	3,634,076	31,021,822	33,705,684
0.5	6,452,520	6,185,955	4,261,836	4,261,836	36,511,538	39,544,692
0.4	8,469,945	8,131,765	5,414,919	5,481,501	42,315,752	45,564,696
0.3	11,796,524	11,352,915	6,886,196	6,962,520	54,931,781	58,524,121
0.2	22,087,957	21,492,219	9,956,201	10,056,323	88,567,223	93,046,528
0.1	32,199,871	31,537,172	16,274,551	16,420,141	122,760,209	128,131,966

Table 6. The average / standard deviation for the number of words (alpha-numeric tokens) in sentences

Score	English-German		English-Romanian		English-Spanish	
	English	German	English	Romanian	English	Spanish
0.9	14.4 / 8.9	14.1 / 8.6	19.3 / 11.0	19.6 / 11.2	12.3 / 8.3	12.6 / 8.7
0.8	17.4 / 9.2	16.8 / 8.8	21.0 / 10.9	21.4 / 11.2	18.2 / 10.7	19.6 / 11.7
0.7	18.4 / 9.8	17.7 / 9.4	21.0 / 10.8	21.3 / 11.1	19.5 / 10.9	21.1 / 12.0
0.6	19.2 / 10.0	18.4 / 9.6	21.1 / 10.8	21.4 / 11.2	19.6 / 10.8	21.3 / 11.9
0.5	20.0 / 10.5	19.2 / 10.0	21.2 / 10.7	21.5 / 11.0	19.9 / 10.9	21.5 / 11.9
0.4	20.5 / 10.5	19.7 / 10.0	21.5 / 10.6	21.7 / 10.9	20.1 / 10.9	21.7 / 11.9
0.3	21.1 / 10.3	20.3 / 9.8	21.7 / 10.7	21.9 / 11.0	20.7 / 10.9	22.0 / 11.8
0.2	23.8 / 10.9	23.1 / 10.8	22.1 / 10.5	22.4 / 10.9	23.0 / 11.9	24.2 / 12.5
0.1	25.2 / 11.4	24.7 / 11.4	23.8 / 11.5	24.0 / 12.0	24.4 / 12.5	25.5 / 13.2

Looking at the average lengths of the sentences (see Table 6), one may notice that the higher the threshold, the shorter the sentence pairs LEXACC finds. This means that the application is more confident when assigning higher scores to shorter sentence pairs. The reader can also deduce that in general, expressing the same statement requires fewer words in German than in English and more words in Spanish than in English, while Romanian requires almost the same number of words as English.

Table 7. Examples of aligned sentences for English-Spanish having scores between 0.1 and 0.9

Score	English-Spanish sentence pair
0.9	The law provides for freedom of assembly and association, and the government generally respected these rights in practice. La ley provee de libertad de asamblea y asociaciones, el gobierno generalmente respeta estos derechos en práctica.
0.8	The rising Swedish exodus was caused by economic, political, and religious conditions affecting particularly the rural population. El creciente éxodo sueco fue causado por condiciones económicas, políticas y religiosas que afectaban particularmente a la población rural.
0.7	After she and her younger brother Andreas began to get successful in skiing - Hanni won the gold medal in slalom at the 1974 World Championships - the family was granted Liechtenstein citizenship. Después ella y su hermano Andreas llegaron a tener éxitos esquiando - Hanni llegó a ser la Campeona Mundial de Slalom en 1974 - a la familia se le concedió la ciudadanía de Liechtenstein.
0.6	Clairemont is a suburban neighborhood in northern San Diego. Clairemont es un barrio localizado en la ciudad de San Diego.
0.5	The origin of the name manganese is complex. El dióxido de manganeso se utiliza como cátodo.
0.4	Although the fossil record of pycnogonids is scant, it is clear that they once possessed a coelom, but it was later lost, and that the group is very old. Los fósiles conocidos de mayor edad son del Devónico, aunque dada su posición sistemática, el grupo debe ser mucho más antiguo.
0.3	Prayer vigils were also held on the second anniversary of the raid in Waterloo and Postville. La redada de Postville fue una redada en una planta empacadora de productos cárnicos en Postville, Iowa.
0.2	Although it sought to avoid entering the war, Spain did make plans for defence of the country. La situación de colaboración con los agentes del Eje, principalmente alemanes, en España era de conocimiento público.
0.1	Three years before events of the game, Dick disappeared from the Hamilton household and has not been seen since. Es asesinado por Dick Hamilton en un ataque de furia, al negar que Alyssa fuese criada como una "Rooder".

Looking at the standard deviations, we see high values which mean that the lengths of the sentences vary a lot and that the normal distributions of the lengths are flat.

The score LEXACC assigns to each extracted pair is an intrinsic measure of the degree of parallelism between the sentences in that pair. Manually analyzing the extracted data, we came to the conclusion that the pairs having scores above 0.4 can be easily considered comparable, while those above 0.6 can be considered parallel. It is up to the user to decide which threshold (s)he wants to use, depending on the personal view of the translation equivalency relation. This can be more rigid, meaning that the user is looking for word-to-word translations, or it can be more permissive, when the user is also satisfied with cross-lingual paraphrases. Table 7 shows English-Spanish sentence pairs with different scores. As the scores decrease, the sentence pairs are less and less reciprocal translations. In the same table, we can see that under the 0.6 value, the degree of parallelism decreases progressively. Yet, one can still find parallel pairs having low value scores assigned. This is probably because LEXACC did not have enough information to assign a higher score. For example, the following sentence pair received a score value of 0.288, which can be considered to be too low:

- The plans were revised just once after construction began, when certain technical difficulties arose.
- Tras el comienzo de las obras el plano aún tuvo que alterarse una vez, para sortear las dificultades técnicas que surgieron.

6 Conclusions

This paper describes a collection of parallel sentences extracted from Wikipedia for three pairs of languages: English-German, English-Romanian and English-Spanish. To do this, we employed LEXACC, a tool for extracting parallel sentences from comparable corpora, developed during the ACCURAT project. Each sentence pair is assigned a score which is a translation similarity measure for the sentences forming the pair. The entire collection of sentence pairs is publically available and can be downloaded from: http://ws.racai.ro:9191/repository/search/?q=Parallel+Wiki. It offers the scientific community almost 7 million comparable sentences, out of which more than 2 million can be safely considered parallel, having a translation similarity score above 0.6. Although several other researchers conducted experiments on extracting parallel sentences from Wikipedia for various language pairs, the volume of their data is, in general, much smaller than in our case. The reason is partly due to initial available Wikipedia data for the considered language pairs, but mainly due to the merits of LEXACC technology[9].

We can conclude that LEXACC is a robust tool which performs well on corpora having Wikipedia's level of comparability and that the whole experiment can be repeated for any other pair of languages in Wikipedia. Certainly, the most productive experiments would be run on well-covered language pairs. Moreover, given the large number of document pairs we had to consider (the entire Wikipedia for three pairs of

[9] http://nlptools.racai.ro/nlptools/index.php?page=lexacc ; see also http://ws.racai.ro:9191/

languages: over 10 Gb of raw text in comparable documents), LEXACC is an efficient tool as it acquired about 2.1 Gb of data (more than 20%) and also a fast tool as the entire running time took a few days over a month (on a fast computer).

A reasonable question that arises refers to the evaluation of the percent of data that can be extracted in this manner from a comparable text. We are convinced that this percent is highly dependent on the level of comparability of the input data. Since this aspect is very hard to be evaluated, the above question can possibly be answered only after conducting much more similar experiments. In the near future, we will evaluate the quality of the extracted data by using it for building translation models that will be tested using our SMT systems.

Acknowledgements

This work has been supported by the METANET4U SPS European project under the Grant Agreement no. 270893 and by the ACCURAT project within the FP7 Program under the Grant Agreement no. 248347.
The authors are also grateful to the anonymous reviewers for their valuable recommendations and comments, which helped them improve the quality of the paper.

References

1. Adafre, S. F., and de Rijke, M.: Finding similar sentences across multiple languages in wikipedia. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, pages 62-69 (2006).
2. Birch, L., Callison-Burch, C., Osborne, M., and Post, M.: The Indic multi-parallel corpus. http://homepages.inf.ed.ac.uk/miles/babel.html (2011).
3. Fellbaum, C.: WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press (1998).
4. Gale, W. A., and Church, K. W.: A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics 19 (1), pages 75-102 (1993).
5. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the 45th Annual Meeting of the ACL Companion Volume Proceedings of the Demo and Poster Sessions, Prague, pages 177-180 (2007).
6. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit Vol. 5 (2005).
7. Mohammadi, M., and Ghasem-Aghaee, N.: Building bilingual parallel corpora based on Wikipedia. In Computer Engineering and Applications (ICCEA), 2010 Second International Conference on, Vol. 2, pages 264-268. IEEE (2010).
8. Och, F.J., and Ney, H.: A systematic comparison of various statistical alignment models. Computational linguistics 29.1, pages 19-51 (2003).
9. Quirk, C., Udupa, R., and Menezes, A.: Generative Models of Noisy Translations with Applications to Parallel Fragment Extraction. In: Proceedings of the MT Summit XI, European Association for Machine Translation, pages 377-384 (2007).

10. Rauf, S., and Schwenk, H.: Parallel sentence generation from comparable corpora for improved SMT. Machine Translation, 25(4), pages 341-375 (2011).
11. Resnik, P., and Smith, N. A.: The web as a parallel corpus. Computational Linguistics, 29(3), pages 349-380 (2003).
12. Smith, J. R., Quirk, C., and Toutanova, K.: 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, pages 403-411. Association for Computational Linguistics (2010).
13. Ştefănescu, D., Ion, R., and Hunsicker, S.: Hybrid parallel sentence mining from comparable corpora. In Proceedings of the 16th Conference of the EAMT, Trento, Italy (2012).
14. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. and Varga, D.: The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pages 2142-2147, Genoa, Italy. ELRA. ISBN 2-9517408-2-4 (2006).
15. Tiedemann, J.: Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC) (2012)
16. Tillmann, C.: A Beam-Search extraction algorithm for comparable data. In Proceedings of ACL, pages 225-228 (2009).
17. Tufiş, D., Ion, R., Bozianu, L., Ceauşu, A. and Ştefănescu, D.: Romanian Wordnet: Current State, New Applications and Prospects. In Proceedings of 4th Global WordNet Conference, GWC-2008, pages 441-452, Szeged, Hungary. University of Szeged, Hungary. ISBN 978-963-482-854-9 (2008).
18. Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D.: RACAI's Linguistic Web Services. In Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008, Marrakech, Morocco (2008).
19. Wu, D.: Aligning a parallel English-Chinese corpus statistically with lexical criteria. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pages 80-87. Association for Computational Linguistics (1994).
20. Yasuda, K., and Sumita, E.: Method for building sentence-aligned corpus from wikipedia. In 2008 AAAI Workshop on Wikipedia and Artificial Intelligence (WikiAI08). (2008).
21. Zhao, B., and Vogel, S.: Adaptive parallel sentences mining from web bilingual news collection. In Proceedings of the 2002 IEEE International Conference on Data Mining, page 745. IEEE Computer Society (2002).