MultiUN: A Multilingual Corpus from United Nation Documents

Andreas Eisele, Yu Chen

German Research Center for Artificial Intelligence (DFKI) GmbH
Saarbrücken, Germany
{Andreas.Eisele, Yu.Chen}@dfki.de

Abstract

This paper describes the acquisition, preparation and properties of a corpus extracted from the official documents of the United Nations (UN). This corpus is available in all 6 official languages of the UN, consisting of around 300 million words per language. We describe the methods we used for crawling, document formatting, and sentence alignment. This corpus also includes a common test set for machine translation. We present the results of a French-Chinese machine translation experiment performed on this corpus.

1. Introduction

The development of machine translation systems, regardless the paradigms behind them, requires a significant amount of knowledge resources in both languages. The essential resource for building a high-quality statistical machine translation (SMT) system is a well-prepared parallel corpus of a reasonable size. Lack of such a corpus is one of the most common reasons of the low translation quality for certain language pairs. There are many other cross-lingual applications relying on parallel corpora, including parallel term extraction, cross-lingual information retrieval, and cross-lingual question answering. In addition to multilingual applications, parallel corpora are also valuable source for propagating linguistic knowledge across languages, which is especially important for morphological (Snyder and Barzilay, 2008) and syntactic analysis (Hwa et al., 2005) for resource poor languages.

As the coverage of language pairs has been extended to include Chinese and Arabic in the EuroMatrixPlus project, we collected parallel texts from official documents through the Official Document System (ODS) of the United Nations in all six official languages, namely Arabic, Chinese, English, French, Russian and Spanish, with a small part of the documents available also in German.

In addition to the processed individual documents, we also provide sentence-aligned bilingual texts of all 21 language pairs. The current release of this MultiUN corpus contains $463,406$ documents encoded in XML, around 1% of them is in German. Over $10,000$ new files have been added to the corpus since previous internal pre-release 6 months ago. The corpus can easily grow when new documents become available in ODS. We hope this corpus can play the role of a bridge from the resources in European languages to the others. The paper explains the automatic acquisition procedure of the corpus and a step further towards its applications in statistical machine translation: a set of SMT systems constructed upon the corpus.

2. Background

In general, the more data is used, the better the translations the system may produce. The size of a parallel corpus directly influences on the translation quality of the SMT system built based on it. Furthermore, the domain and the time range of the corpus also have an effect on the SMT system, since the the use of languages varies significantly under different conditions. Hence, given the domain of the corpus, it is always useful to have a corpus from a source providing up-to-date new documents/texts such as EuroParl (Koehn, 2005).

While early parallel texts usually involve only two languages, one of which is English in many cases, many corpora developed in recently years contain versions of texts in more languages. Such multilingual aligned texts provide a more compact way for storage. A parallel corpus in N languages is equivalent to at least $N(N-1)/2$ bilingual corpora. As a matter of fact, multilingual corpora carry far more information than the union of the corresponding bilingual corpora. The correspondences among multiple languages are able to reveal hidden knowledge that cannot be easily inferred from any single correspondence. One instance of utilizing such corpora is triangulation, either through the unions of multiple translation correspondences (Cohn and Lapata, 2007; Och and Ney, 2001; Kumar et al., 2007; Simard, 1999) or through the intersections (Chen et al., 2008; Chen et al., 2009). To our knowledge, this type of corpora mostly exists for European languages, such as Europarl (11 European langugaes), UMC (Klyueva and Bojar, 2008) (English-Czech-Russian), UN Parallel text (Graff, 1994) (English-French-Spanish) and JRC-Acquis (Steinberger et al., 2006) (23 European languages) .

Multilingual corpora of European languages, such as EuroParl and JRC-Acquis, were the essential materials to produce as many as 462 machine translation systems (Koehn et al., 2009). Although multilingual data acquired from UN documents has already been used/discussed by many researchers (Eisele, 2006; Kumar et al., 2007) and a limited subset of the official documents is available (Rafalovitch and Dale, 2009), there is no corpus comparable to the one we present here.

3. Corpus collection

We describe here briefly the acquisition procedure of the MultiUN corpus from the ODS of the United Nations. The documents we collected are in public domain according to the Administrative Instruction

(ST/AI/189/Add.9/Rev.2) from the United Nations (United Nations Secretariat, 1987).

Crawling We crawled a set of documents from the ODS website of the United Nations, where most of the documents are encoded in Microsoft Word DOC format. The bulk of the data we obtained is from the years 2000 up to 2009, and before document alignment and cleaning, we had more than a million documents per complete year.

Preprocessing We converted the documents collected from the ODS to XML. In order to preserve high quality texts, we discarded certain contents in the original documents, including the pictures, the tables and some style markers. Besides, the converted files also contain the following meta-information regarding the original files:

- a file ID, which is unique for each individual file,
- the language with which the document is labeled in the ODS,
- the publication date of the document, and
- a document *symbol*, a unique identifier consisting of numbers and letters for a United Nations document.

A *symbol* in ODS is a unique identifier consisting of numbers and letters for a United Nations document. The first component in a symbol usually indicates the parent organ issuing the document or to which the document is being submitted and the rest of the components may reflect subsidiary bodies, the nature of the document or modifications to the original text, which appear as suffixes. For instance, the symbol "A/CONF.157/PC/63/Add.4" refers to document No. 63 from the Preparatory Committee of the World Conference on Human Rights in General Assembly. The symbol is shared by all language versions of a document. In other words, we can easily identify parallel documents by the symbols. It is possible for one document to be related to several dates. We only preserved the final publication date to avoid confusion.

Selection and cleaning Not all the extracted contents are fully suitable for direct use in natural language processing applications. There are several types of meta-information in these documents beyond the text. Even though it is always desirable to collect as many texts as possible, it is even more important to ensure the quality of the texts, which is crucial for the systems built from the texts. We are thereby strict on selection of the documents. Most documents before 1993 are only scanned pictures. Moreover, many files before 2000 are formatted in various encodings other than unicode depending on the languages, which became the major obstacle for the extraction of useful texts. Thus, we processed documents only starting from year 2000 due to this irregularity. As for the documents we kept, we filtered out all out-of-range characters as well as noisy lines, for example, lines with only URL links.

The language of each individual document is another criterion for our selection. We observed that the language information provided in the ODS is not always reliable. Including texts in a different language will certainly introduce noise to the corpus. The size of the texts in a language different from the one given for the file varies from one line to the complete document. To find out the documents inconsistent with their language labels, we applied *mguesser*(Barkov, 2008), a C implementation of the N-gram based text classification algorithm *TextCat* (Cavnar and Trenkle, 1994), to all the documents. We manually selected 10 manually verified files, roughly 500Kb, to build a customized character-based language model for each language. A file is removed from our collection if it is identified as different from the given language. Meanwhile, the annexes in the documents, mostly duplicated rather than translated across the languages, are the main source of such noisy texts. We discarded any documents containing annexes without further language identification. It is also possible to introduce more fine grained language identification to preserve more texts in the future.

Formating After the documents are selected and cleaned, there are a few more steps to facilitate the use of the corpus. Firstly, the UN official documents only provide paragraph boundaries. However, most machine translation tasks consider sentences as the basic units, as the complexity of both system training and translations increases significantly with longer units. We conducted sentence splitting to the extracted texts to follow this tradition. The Chinese texts are split with simple regular expressions, while the others are processed using the sentence tokenization module from NLTK toolkit (Bird and Loper, 2004). However, the paragraph boundaries are preserved in the documents.

Secondly, the files are originally assigned by ODS with names not related to the documents' properties in any obvious way. We renamed the files according to the document symbol and the language. Since some characters in the document symbols are not allowed to appear in file names in certain operating systems, we abbreviate the symbols to meet the requirements while maintaining the unambiguity of the symbols. For instance, the file originally labeled with the file id "*N0831582*" is now named "*S_AGENDA_5874-ru*", where "*S_AGENDA_5874*" indicates the document symbol "*S/AGENDA/5874*" and "*ru*" refers to Russian.

Finally, we introduced several additional XML tags to indicate special phrases. Only URLs and emails are identified and tagged so far for the current release, but it is clearly possible to include further tags, especially for named entities like person names, locations, organization names, etc.

Sentence alignment As the last step, the sentences are automatically aligned with *hunalign* (Varga et al., 2005), which takes tokenized sentence-segmented texts in two languages and outputs a sequence of bilingual sentence pairs. Without dictionaries for all languages, we chose the two-pass option that first aligns based on sentence-length information, builds an automatic dictionary based on this alignment and realigns the text in a second pass, using the automatic dictionary. The paragraph boundaries are also considered for the alignments. However, the changes of

sentence order are not handled. Aligned texts are prepared bilingually for all 21 language pairs between the 7 languages. We plan to deliver multilingual alignments in the near future.

Common Test Set We plan to renew the corpus every half a year. Since the documents usually become available in 6 months after its original publication date, it is only necessary to process documents for the last 12 months in each update. Furthermore, we reserved the documents in the last 6 months for testing and comparisons, which is going to be included in the training set in the next update. This not only allows us to increase the size of the training set but also help to compare the MT systems on a up-to-date test set.

4. Availability of the corpus

We made the corpus available to the research community through the web site of the EuroMatrixPlus project [1]. We hope that free access to this parallel corpus will be beneficial for extending the coverage of publicly available MT systems to additional languages in similar ways as Europarl and the Acquis Communautaire were crucial for research and development of MT systems for EU languages in the recent past.

5. Property of the corpus

The current version of the corpus consists of documents from January 2000 to September 2009. Meanwhile, the documents from October 2009 to March 2010 are included as testing material.

Table 1 lists all sections of the corpus by language. The corpus consists of $463,406$ documents, $80,931,645$ sentences in total. There are 326 million tokens on average for five of the six official languages.

We disregarded Chinese for all word counts as there are many distinct standards for Chinese word segmentation. Hence, we only report the counts of Chinese characters in the corpus here.

More statistics of the aligned parallel texts for each language pairs are given in Table 2, 3 and 4.

6. Application in MT

As the first step towards applying the corpus in MT, we use this corpus to construct SMT systems with the Moses toolkit (Koehn et al., 2007). Translation models were trained using the complete set of parallel texts, while 5-gram language models were built with only the target side of the aligned texts. We randomly extracted $3,000$ sentences from the reserved portion of the corpus, $1,000$ of which were used for training the feature weights of the decoder and the rest was for testing.

Using the Moses toolkit (Koehn et al., 2007), we followed a standard routine to build SMT systems for four translations directions between three languages: Spanish to Chinese, Chinese to Spanish, French to Chinese and Chinese to French. We segmented the Chinese texts with Stanford Chinese Word Segmenter (Tseng et al., 2005). The phrase-based translation models are trained on sentences with less than 80 tokens.

The systems also include 5-gram language models trained on the target side of corresponding parallel texts using SRILM (Stolcke, 2002). The feature weights required by the Moses decoder are further determined with minimum error rate training (MERT) (Och, 2003) by optimizing BLEU scores on the 1,000-sentence development set. The test sets were translated by the resulting systems and thus used to evaluate the systems in terms of BLEU scores (Papineni et al., 2001), as shown in Table 5

Language pairs	Development set	Test set
Spanish-Chinese	33.25	31.35
Chinese-Spanish	40.65	39.08
French-Chinese	29.40	29.94
Chinese-French	34.85	34.66

Table 5: BLEU scores of the SMT systems

Figure 1 shows an example translation produced by our Chinese-French system together with translation of the same sentence provided by an online MT engine. Our system is able to correctly translate the organization names and the dates, even though we did not include preprocessing such as named entity recognition into the system. It is mostly likely because of the large amount of in-domain training data.

7. Conclusion

We have presented MultiUN version 1, the new release of the multilingual parallel corpus extracted from official documents of the United Nations. Using the corpus, we were able to construct SMT systems for several language pairs that were not often studied before and the translation results are satisfactory given the amount of efforts required to build the systems. However, the released corpus may still contain various errors and flaws. We plan to further improve our processing techniques and provide more features in further versions.

We believe MultiUN corpus is a very useful resource for MT developers. The corpus not only allows us to build SMT systems for the 21 translation directions, but also provides many potential bridges between European languages that are included in existing multilingual parallel corpora to other languages, such as Arabic and Chinese. We hope the release of this corpus will further boost the research in the field.

Acknowledgements

This work was supported by the European Community through the EuroMatrix Plus project (ICT-231720) funded under the Seventh Framework Programme for Research and Technological Development.

[1] http://www.euromatrixplus.eu/downloads

[2] Only number of characters are counted for Chinese texts.

Language	English	French	Spanish	Arabic	Russian	Chinese	German
Documents	96240	85651	70509	65156	77061	65022	3763
Sentences	17098695	14805529	13052875	11050313	13852535	10839473	232225
Words[2]	385894793	377242310	352460926	237412090	278606813	756108566	5848668

Table 1: Sizes of monolingual data

	fr	es	ar	ru	zh	de
en	96240	68314	63257	74053	62815	3643
fr		68014	63193	73973	62738	3632
es			63241	64230	62707	3632
ar				63194	63031	3677
ru					62842	3635
zh						3886

Table 2: Number of document pairs for the language pairs

	fr	es	ar	ru	zh	de
en	12317630	10663070	8206568	6166942	8829060	156507
fr		11062423	8841529	8244246	8727150	153177
es			8794485	6873585	8788366	150874
ar				9045096	7581084	144408
ru					4778903	120836
zh						153815

Table 3: Number of sentence pairs for the language pairs

	fr	es	ar	ru	zh	de
en	301878068	267327033	214681635	163816832	220423478	4770788
	337798816	311593654	180759040	97272528	629509919	4626954
fr		315786275	255119236	236294787	243833077	5332166
		328728803	195015089	154383105	627949668	4583826
es			262276003	209978115	252782013	5511988
			193169242	128538959	621772385	4494109
ar				200786093	162186078	3976850
				196836622	574219433	4384192
ru					68484167	3089380
					392289340	3758626
zh						13109304
						4525561

Table 4: Number of words (L1:L2) for the language pairs [2]

8. References

Alexander Barkov. 2008. mguesser. http://www.mnogosearch.org/guesser/.

Steven Bird and Edward Loper. 2004. NLTK: the natural language toolkit. In *ACL 2004 on Interactive poster and demonstration sessions*, page 31, Morristown, NJ, USA. Association for Computational Linguistics.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175.

Yu Chen, Andreas Eisele, and Martin Kay. 2008. Improving Statistical Machine Translation Efficiency by Triangulation. In *the 6th International Conference on Language Resources and Evaluation (LREC '08)*, May.

Yu Chen, Martin Kay, and Andreas Eisele. 2009. Intersecting multilingual data for faster and better statistical translations. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 128–136, Boulder, Colorado, June. Association for Computational Linguistics.

Trevor Cohn and Mirella Lapata. 2007. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In *the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech, June.

Andreas Eisele. 2006. Parallel Corpora and Phrase-Based Statistical Machine Translation for New Language Pairs via Multiple Intermediaries. In *LREC 2006*, Genoa, Italy, May.

David Graff. 1994. UN Parallel Text (Complete). Linguistic Data Consortium, Philadelphia.

[4]

Source:	国民议会第二次常会于5月4日至22日举行。
Reference:	La deuxième session spéciale de l'Assemblée nationale a eu lieu du 4 au 22 mai.
In-house:	L'Assemblèe nationale à la deuxième réunion ordinaire du 4 au 22 mai .
Online:	La deuxième session ordinaire de l'Assemblée nationale le 4 Mai 22 a été tenue.

Figure 1: Sample translation from the Chinese-French SMT system compared to an online MT engine

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clare Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325, September.

Natalia Klyueva and Ondřej Bojar. 2008. UMC 0.1: Czech-Russian-English Multilingual Corpus. In *International Conference Corpus Linguistics*, pages 188–195, October.

Philipp Koehn, Hieu Hoang, Alexandra Birch Mayne, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbs. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of Annual meeting of the Association for Computation Linguistics (acl), demonstration session*, pages 177–180, Prague, Czech, June.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. 462 Machine Translation Systems for Europe. In *Machine Translation Summit XII*, pages 65–72, August.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit 2005*.

Shankar Kumar, Franz Josef Och, and Wolfgang Macherey. 2007. Improving word alignment with bridge languages. In *the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 42–50, Prague, Czech.

Franz Josef Och and Hermann Ney. 2001. Statistical multi-source translation. In *MT Summit VIII*, Santiago de Compostela, Spain.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Alexandre Rafalovitch and Robert Dale. 2009. United nations general assembly resolutions: A six-language parallel corpus. In *Proceedings of the MT Summit XII*, pages 292–299. International Association of Machine Translation, August.

Michel Simard. 1999. Text-translation alignment: Three languages are better than two. In *EMNLP/VLC-99*, College Park, MD, June.

Benjamin Snyder and Regina Barzilay. 2008. Cross-lingual propagation for morphological analysis. In *AAAI'08: Proceedings of the 23rd national conference on Artificial intelligence*, pages 848–854. AAAI Press.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Daniel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *LREC 2006*, Sep.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *the 7th International Conference on Spoken Language Processing (ICSLP) 2002*, Denver, Colorado.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter. In *Fourth SIGHAN Workshop on Chinese Language Processing*.

United Nations Secretariat. 1987. Regulation for the control and limitation of documentation. Administrative Instruction (ST/AI/189/Add.9/Rev.2), 17 September.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *RANLP*, pages 560–596, Borovets, Bulgaria.