MultiUN v2: UN Documents with Multilingual Alignments

Yu Chen†, Andreas Eisele‡

†LT-Lab DFKI GmbH
D-66123 Saarbrücken, Germany
Yu.Chen@dfki.de
‡Directorate-General for Translation (DGT)
L-2920 Luxembourg
Andreas.Eisele@ec.europa.eu

Abstract

MultiUN is a multilingual parallel corpus extracted from the official documents of the United Nations. It is available in the six official languages of the UN and a small portion of it is also available in German. This paper presents a major update on the first public version of the corpus released in 2010. This version 2 consists of over $513,091$ documents, including around 9% of new documents retrieved from the United Nations official document system. Compared to the first release, we applied several modifications to the corpus preparation method. In this paper, we describe the methods we used for processing the UN documents and aligning the sentences. The most significant improvement compared to the previous release is the newly added multilingual sentence alignment information. The alignment information is encoded together with the text in XML instead of additional files. Our representation of the sentence alignment allows quick construction of aligned texts parallel in arbitrary number of languages, which is essential for building machine translation systems.

Keywords: United Nations, parallel corpus, multilingual, machine translation, crosslingual

1. Introduction

Parallel corpora have become essential resources for many natural language processing (NLP) applications. The quality of the parallel corpus used as training data is extremely critical for building a high quality statistical machine translation (SMT) system. Many rule-based machine translation (RBMT) systems also consist of components that are constructed based on parallel texts. Apart from machine translation, parallel corpora play an important role in other cross-lingual applications, such as cross-lingual information retrieval.

In recent years a growing number of parallel corpora are constructed for more than two languages at the same time as they are derived from text collections translated to multiple languages (Koehn, 2005; Klyueva and Bojar, 2008; Steinberger et al., 2006; Tiedemann, 2009). Such multilingual corpora not only store pairwise translations more efficiently, but also supply more correspondence information among the languages. Meanwhile, the continuously evolving topics and styles of the written texts have noticeable effects on NLP applications. Besides, the performance of many methods, especially statistical ones, relies on the amount of training materials. Hence, our aim for MultiUN is to construct a multilingual parallel corpus that grows with up-to-date texts continuously.

MultiUN is a multilingual corpus extracted from the official documents of the United Nations (UN) available in 6 official UN languages (Eisele and Chen, 2010). After its first release, the corpus has been included as training data in several evaluation events on machine translation (Callison-Burch et al., 2010; Callison-Burch et al., 2011; Federico et al., 2011). This release of the corpus extends the previous version with additional two years of documents. We refine the cleaning procedure and introduce new annotations to the corpus. The main contribution of this update is that this version includes *multilingual* sentence alignments that were unavailable in the previous releases. We present the alignments of MultiUN as embedded annotations directly wrapped around the texts. An extraction script is provided together with the corpus for extracting texts sentence-aligned for an arbitrary number of languages.

2. Previous work

Many multilingual corpora have been developed in recent years. A majority of such corpora exists only for a few European languages, such as Europarl (Koehn, 2005), UMC (Klyueva and Bojar, 2008), UN Parallel Text (Graff, 1994) and JRC-Acquis (Steinberger et al., 2006).

Among the existing multilingual parallel corpora, there are several different ways to supply the sentence alignment. One way is to include a sentence alignment tool in the corpus, e.g. EuroParl, so the user can extract sentence aligned texts on demand. Another way is to remove any unaligned sentences and present the aligned sentence pairs, e.g. UMC and OPUS (Tiedemann, 2009). As a result, many sentences need to be duplicated several times in the alignment files. Alternatively, the alignment files in Acquis (Steinberger et al., 2006) only include pointers to the text files. Nevertheless, all the current multilingual corpora only supply bilingual alignments. No multilingual alignments are possible to our knowledge.

3. Corpus collection

This section briefly describes the acquisition procedure of the MultiUN corpus from the Official Document System (ODS) of the United Nations. The documents we collected are in public domain according

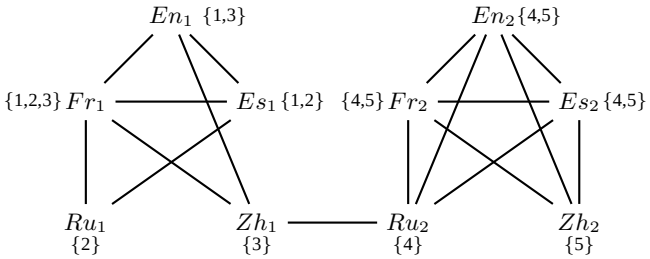

Figure 1: Pairwise alignments

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

Crawling We collected documents from year 2000 up to 2011 from the ODS website of the United Nations. A document could have been released multiple times in the system. Only the latest version is included in the corpus.

Preprocessing The original files are in Microsoft Word format. We first extract only the plain texts from the collected files and remove all the footnotes, figures, graphics, tables, hyper links and many other non-text contents.

The extracted texts are then split into sentences. The Chinese sentences are identified with regular expressions. For the other 6 languages, we apply a language independent unsupervised approach to disambiguate the sentence boundaries from abbreviations (Kiss and Strunk, 2006). During sentence segmentation, paragraph boundaries are preserved.

On top of the segmented texts, we construct structured XML files with information indicating the origins of the files, including the file ID's, the languages, the publication dates and the so-called "document symbols". The document symbol is unique for a document regardless of the version or the language of the files. Hence, we consider the symbols as the indicator of the parallel documents.

Selection and cleaning In order to ensure the quality of the texts in corpus, we are fairly strict on document selection.

First, any documents published before 2000 are excluded for further processing due to various types of technical issues. The documents from the last 6 months are again reserved for testing and comparison with the systems built on previous release. The current test set is going to be included in the next update.

Second, we send each individual document to a language identification software *mguesser* (Barkov, 2008) trained on manually verified documents. If the identification result is inconsistent with the language indicated in the ODS, the document would be discarded.

Finally, a document will be removed from the collection if the ratio of the noisy texts such as illegal characters, foreign words, etc. is too high according to a rule set. The rule set is being updated accumulatively, also based on feedback from the users.

4. Multilingual sentence alignment

The multilingual sentence alignment of MultiUN starts with pairwise alignments. The sentences in a pair of parallel documents are first aligned based on their lengths (Gale and Church, 1991). Based on a dictionary generated from this alignment, the sentences are aligned again to form the final alignment. We align the texts bilingually in this way for all 21 language pairs using *hunalign* (Varga et al., 2005). We do not try to detect or handle reordering of sentences between the translations. All pairwise alignments are computed for each group of corresponding documents.

There are at least two straightforward methods to construct multilingual alignments from pairwise alignment results. One way is to union a minimum number of pairwise alignments that covers all languages, which usually leads to larger alignment units, higher alignment coverage, but also most likely lower precisions. Another way is to intersect all given pairwise alignments. In this case, many alignment links are removed from the pairwise alignments.

Figure 1 illustrates the pairwise alignments between the 10 sentences in 5 languages. The sentences are identified by their language (En, Fr, Es, Ru or Zh) and index (1 or 2) in a document. As the sentence Zh_1 and Ru_2 are aligned, all the other sentences are connected through this link. Thus, the union method takes the whole set of sentences as an alignment group, while the intersect method discards all the links.

Both methods rely on the assumption of transitivity of sentence alignments, that is, if sentence a corresponds to sentence b, and if b itself corresponds to a third sentence c, then a also corresponds to c. In practice, this assumption does not always hold for multilingual documents as the segments of translations are not necessarily consistent with the sentence boundaries. However, it is still clear that indirect alignments through other languages are able to imply the possible direct alignments between two languages. That is, the more languages in which common translations exist for the two sentences, the higher the chance of the two sentences being translations of each other.

Our approach aims at improving the alignment accuracy while preserving the information generated during the pairwise alignments. The method is fairly simple, given a complete graph of pairwise alignments. For each (pairwise) alignment link, we first examine whether the two sentences are connected through a sentence in any other language. If not, we check whether both sentences are aligned to some sentences in the same third language. If the two sentences are aligned to different sentences in the same language, the

alignment between these two sentences are considered *inconsistent* with the other alignments. We delete such inconsistent alignments from the graph. The alignment connecting Zh_1 and Ru_2 in Figure 1 should be removed.

After this validation step, the sentences are grouped by the alignments remaining in the graph. A set of maximally fully connected sentences is marked as one group. No sentence outside a group should be aligned to all sentences in that group, but one sentence may belong to multiple groups. We can extract multilingual alignments of an arbitrary number of languages simply by traversing the groups. There are 5 groups in the example discussed above. Each node in the graph is marked with the groups it belongs to.

5. Property of the corpus

The current version of the corpus consists of documents from January 2000 to June 2010. The documents from later on are included as testing material. We describe the format of MultiUN and present a few statistics of this corpus in this section.

5.1. Corpus format

We introduce the sentence alignment information as an additional attribute in the XML documents.

The upper part of Figure 2 shows a few segments of an English document in MultiUN v2. It was published in 2009 and the document symbol is "SAICM/ICCM.2/6". The original file ID was "K0950702" and it was last updated in February 2009. This document is available for all six official languages, but not German.

Apart from the paragraph and sentence index (*n*), each sentence is assigned with an alignment list (*aligned*). The list includes all the alignment points that are related to the corresponding sentences. In other words, any sentences that are linked to the same alignment points are aligned as a group of parallel sentences. The corresponding lines in the other version of the same document are given in the lower part of Figure 2. The indices start with '1'. '0' in the *aligned* field indicates this sentence was not aligned to any sentences in other languages.

5.2. Statistics

The basic characteristics of this version of the corpus are listed in Table 1 for each language. Although the filtering rules are more strict for selecting the documents, the current version still consists of 9% new documents that do not exist in the previous release. These new documents added around 5% new sentences to the corpus as we have filtered out more noisy sentences than before. Table 2 shows the number of aligned documents and sentence pairs for each language pair. The addition to the bilingual alignments is consistent with the overall increase.

We measured the coverage of the multilingual sentence alignment using the average ratio of the number of aligned sentence to the overall number of sentences. Table 3 lists the coverage of the multilingual alignments for different numbers of languages involved.

Languages	Coverage ($\frac{aligned\ sentences}{all\ sentences}$)
2	0.98479
3	0.65838
4	0.53123
5	0.44057
6	0.37330

Table 3: Sentence coverage of multilingual alignments

6. Availability of the corpus

This version is available to the research community through the web site of the EuroMatrixPlus project[1] in the same manner as the previous releases. We hope that free access to this parallel corpus, especially the addition of multilingual sentence alignments, will not only be beneficial for research of machine translations between the seven languages in this corpus but also serve as a connection for the previously existing parallel corpora to facilitate development of MT systems of many language pairs for which no direct parallel corpus is available.

7. Conclusion

We presented the latest release of MultiUN corpus that provides multilingual sentence alignments along with around 10% recently collected documents. The multilingual sentence alignments are constructed based on all possible pairwise alignments. We applied simple heuristics to identify the possible errors in pairwise alignment without sacrificing the overall coverage of multilingual sentence alignments. As a result, nearly 40% of the sentences in documents that are parallel in all 6 languages are aligned.

We only consider the bilingual alignments with high confidence scores for multilingual alignments. It should be useful to also take the bilingual alignment scores into account. Besides, we could benefit more from the indirect alignments that we used for validation by searching for missing alignment links. Furthermore, it is no doubt necessary in the future to verify the effects of multilingual alignments on machine translation systems.

Acknowledgements

The first author was supported by the European Community through the EuroMatrix Plus project (ICT-231720) funded under the Seventh Framework Programme for Research and Technological Development. We thank Christian Federmann for maintaining the website. Many thanks to other colleagues from our lab and from the EuroMatrix consortium for the inspiration and testing. We apologize for some overlap with the material presented in (Eisele and Chen, 2010).

8. References

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

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010.

[1] http://www.euromatrixplus.net/multi-un

```
<?xml version="1.0" encoding="UTF-8"?>
<DOC lang="English" n="K0950702" id="SAICM/ICCM.2/6" date="2009/02/18">
  <text>
    <body>
      ...
      <p n="4">
          <s n="4" aligned="6"> Agenda item 4 (e) </s>
      </p>
      <p n="5">
          <s n="5" aligned="7"> Implementation of the Strategic Approach to International Chemicals
Management: financial and technical resources for implementation </s>
      </p>
      <p n="6">
          <s n="6" aligned="8"> Summary and commentary on submissions received from stakeholders in
response to the questionnaires on financial arrangements for the Strategic Approach to International
Chemicals Management </s>
      </p>
      <p n="7">
          <s n="7" aligned="10,11"> Note by the Secretariat </s>
      </p>
      ...
    </body>
  </text>
</DOC>
```

```
French:
      <s n="4" aligned="6"> Point 4 e) de l&apos;ordre du jour* </s>
      <s n="5" aligned="7"> Mise en œuvre de l&apos;Approche stratégique de la gestion
internationale des produits chimiques : ressources financières et techniques pour la mise
en œuvre </s>
      <s n="6" aligned="8"> Résumé et observations formulées sur les communications reçues
des parties prenantes en réponse aux questionnaires sur les dispositions financières
applicables à l&apos;Approche stratégique de la gestion internationale des produits
chimiques </s>
      <s n="7" aligned="10,11"> Note du secrétariat </s>
```

```
Spanish:
      <s n="4" aligned="6"> Tema 4 e) del programa provisional* </s>
      <s n="5" aligned="7"> Aplicación del Enfoque Estratégico para la Gestión de Productos
Químicos a Nivel Internacional: recursos financieros y técnicos para la aplicación </s>
      <s n="6" aligned="8,9"> Resumen y comentarios sobre las observaciones presentadas por
los interesados directos en respuesta a los cuestionarios sobre los arreglos financieros
para el Enfoque Estratégico para la Gestión de Productos Químicos a Nivel Internacional </
s>
      <s n="7" aligned="10,11"> Nota de la Secretaría </s>
```

```
Arabic:
      <s n="1" aligned="0"> :تنفيذ النهج الاستراتيجي للإدارة الدولية للمواد الكيميائية
الموارد المالية والتقنية اللازمة للتنفيذ </s>
      <s n="2" aligned="8"> موجز وتعليق على الطلبات المقدمة من أصحاب المصلحة استجابة
للاستبيان الخاص بالترتيبات المالية للنهج الاستراتيجي للإدارة الدولية للمواد الكيميائية </s>
      <s n="3" aligned="10,11"> مذكرة من الأمانة </s>
      <s n="4" aligned="13,14"> موجز تنفيذي </s>
```

```
Russian:
      <s n="4" aligned="6"> Пункт 4 e) повестки дня* </s>
      <s n="5" aligned="0"> Осуществление Стратегического подхода к Международному
регулированию химических веществ: финансовые и технические ресурсы, необходимые для
осуществления </s>
      <s n="6" aligned="9"> Резюме и замечания относительно сообщений, полученных от
субъектов деятельности в ответ на вопросники о финансовых договоренностях, касающихся
Стратегического подхода к международному регулированию химических веществ </s>
      <s n="7" aligned="10"> Записка секретариата </s>
```

```
Chinese:
      <s n="6" aligned="0"> 实施工作的财政和技术资源 </s>
      <s n="7" aligned="0"> 关于所收到的利益攸关方回复有关国际化学品管理战略方针 </s>
      <s n="8" aligned="0"> 财政安排的调查问卷的来文的摘要和评注 </s>
      <s n="9" aligned="11"> 秘书处的说明 </s>
```

Figure 2: Sections of a document (SAICM/ICCM.2/6) in XML format with alignment information

Language	English	French	Spanish	Arabic	Russian	Chinese	German
Documents	104469	94890	78747	72469	85407	72742	4367
Sentences	17871877	15600899	14022595	11779207	14743388	11551950	256074
Words	401638558	405314181	376927116	251283550	288205989	520141860	6288600

Table 1: Sizes of monolingual data

	fr	es	ar	ru	zh	de
en	94890	76694	70794	82784	70242	4082
	12886831	11196913	8554061	6427032	9188441	167624
fr		76445	70724	82728	70201	4074
		11639293	9207063	8601324	9093059	164299
es			70982	72322	70439	4078
			9281290	7260492	9291593	162702
ar				71085	70690	4121
				9570688	8010458	154916
ru					70537	4079
					5036874	129610
zh						4329
						165307

Table 2: Number of pairwise aligned documents and sentence pairs

Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, July. Association for Computational Linguistics. Revised August 2010.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland, July. Association for Computational Linguistics.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from United Nation documents. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Marcello Federico, Luisa Bentivogli, Michael Paul, and Sebastian Stüker. 2011. Overview of the iwslt 2011 evaluation campaign. In *International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, USA, 12.

William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL '91, pages 177–184, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32:485–525, December.

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

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.

Jörg Tiedemann. 2009. News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In Nicolas Nicolov, Galia Angelova, and Ruslan Mitkov, editors, *Recent Advances in Natural Language Processing V*, volume 309 of *Current Issues in Linguistic Theory*, pages 227–248. John Benjamins, Amsterdam & Philadelphia.

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.