MulTed: a multilingual aligned and tagged parallel corpus

Imad Zeroual and Abdelhak Lakhouaja
Faculty of Sciences, Department of Computer Sciences, Mohamed First University, Oujda, Morocco

Received 5 June 2018
Revised 13 November 2018
Accepted 13 December 2018

Abstract

Recently, more data-driven approaches are demanding multilingual parallel resources primarily in the cross-language studies. To meet these demands, building multilingual parallel corpora are becoming the focus of many Natural Language Processing (NLP) scientific groups. Unlike monolingual corpora, the number of available multilingual parallel corpora is limited. In this paper, the MulTed, a corpus of subtitles extracted from TEDx talks is introduced. It is multilingual, Part of Speech (PoS) tagged, and bilingually sentence-aligned with English as a pivot language. This corpus is designed for many NLP applications, where the sentence-alignment, the PoS tagging, and the size of corpora are influential such as statistical machine translation, language recognition, and bilingual dictionary generation. Currently, the corpus has subtitles that cover 1100 talks available in over 100 languages. The subtitles are classified based on a variety of topics such as Business, Education, and Sport. Regarding the PoS tagging, the Treetagger, a language-independent PoS tagger, is used; then, to make the PoS tagging maximally useful, a mapping process to a universal common tagset is performed. Finally, we believe that making the MulTed corpus available for a public use can be a significant contribution to the literature of NLP and corpus linguistics, especially for under-resourced languages.

Keywords Multilingual parallel corpora, Natural Language Processing, Linguistic resources, Sentence alignment, Treetagger

Paper type Original Article

1. Introduction

Given their importance, the demand for multilingual parallel resources is increasing primarily for those covering under-resourced languages. However, the problem of building a balanced mix of multilingual texts in sufficient quantities and with a high-quality of translation becomes ever more central. This bottleneck becomes quite prohibitive when any further processing, such as sentence-alignment or Part of Speech (PoS) tagging, are to be involved. Despite the difficulty of building such corpora, they are very valuable for many Natural Language Processing (NLP) applications given that the progress in most of these applications is driven by available data [1]. Statistical Machine Translation (SMT) is one of these applications that have achieved significant impact with the help of such corpora

Publishers note: The publisher wishes to inform readers that the article "MulTed: A multilingual aligned and tagged parallel corpus" was originally published by the previous publisher of Applied Computing and Informatics and the pagination of this article has been subsequently changed. There has been no change to the content of the article. This change was necessary for the journal to transition from the previous publisher to the new one. The publisher sincerely apologises for any inconvenience caused. To access and cite this article, please use Zeroual, I., Lakjouaja, A. (2020), "MulTed: A multilingual aligned and tagged parallel corpus", New England Journal of Entrepreneurship, Vol. ahead-of-print No. ahead-of-print. https://10.1016/j.aci.2018.12.003. The original publication date for this paper was 14/12/2018.

Applied Computing and Informatics
Emerald Publishing Limited
2210-8327
DOI 10.1016/j.aci.2018.12.003

especially if they are aligned at the level of words as well as sentences [2]. The main challenge for the success of the translation process is the access to high-quality training data. This can ease the task for translators, by providing an initial translation, which can be later post-edited.

PoS tagging is a basic task in NLP and corpus linguistics. Knowing that different ambiguity patterns are likely to occur in different places across languages, combining information from many languages creates a clearer picture of each [3]. When a parallel corpus is available, a cross-lingual PoS tagging can be used to assess the effectiveness of cross-linguistic projection of morphological features to an under-specified target language [4].

In contrast to monolingual corpora, only few parallel corpora are available. In addition, their sources cover a restricted range of text types such as legislation, administration, and technical documentation. Furthermore, they are limited to high-density languages such as English and the official European languages.

Taking advantage of the growth of online databases of videos subtitles from TED talk's events, the aim of the work presented in this paper is to present a new multilingual aligned and PoS-tagged parallel corpus called MulTed. The current version of the MulTed corpus includes subtitles of 1100 talks available in over 100 languages. The corpus comprises 30,000+ subtitles that contain 7.6 million aligned sentences with altogether over 46 million words. We sentence-align the entire corpus considering English as a pivot language, i.e., the alignment is done between English and the other languages. Regarding the annotation process, a language-independent tagger called Treetagger [5] and a universal common tagset [6] are implemented. Finally, the subtitles are classified manually into 11 categories based on the variety of TED topics.

In addition to the introduction, the remainder of this paper is arranged into eight main sections. In Section 2, we provide background information about related works, mentioning some relevant parallel corpora. The compilation and filtering tasks of the collected data are described in Section 3. In Section 4, we explore the sentence alignment method used in this work. We exhibit the PoS tagging process performed and the tagset adopted in Section 5. In Section 6, we present statistical information of the current version of MulTed corpus. A sample of the corpus format is given in Section 7. In Section 8, we discuss the advantages and drawbacks of relevant parallel corpora including the MulTed corpus. Finally, we conclude this paper in Section 9 with future enhancements for the next versions.

2. State of the art

In this section, we present some relevant bilingual and multilingual parallel corpora. Since their construction is expensive in terms of time and effort, only few corpora are freely available.

2.1 Bilingual parallel corpora

In last decade, a range of bilingual parallel corpora has been established especially those including the English language. For instance:

- The **CzEng** corpus [7] is a Czech-English parallel corpus and freely available for non-commercial research or educational purposes. Several features have been included in the last release such as morphological tags, surface syntactic, and automatic co-reference links. Most sources of this corpus are books, European Union (EU) Legislation, and Movie Subtitles. By increasing its size, the corpus reached 15 million sentence pairs (about 200 million tokens per language).
- The **Scielo** corpus [8] is a freely available parallel corpus of scientific publications for the biomedical domain (biological sciences and health sciences). The corpus data have

been retrieved from the Scielo database. The corpus is available for three language pairs: Portuguese-English (about 86,000 documents in total), Spanish-English (about 95,000 documents) and French-English (about 2000 documents).

- The **FAPESP** corpus, is constructed in two language pairs, Portuguese-English and Portuguese-Spanish, based on scientific news texts. These texts have been crawled automatically from the online multilingual Brazilian magazine (Pesquisa FAPESP). The corpus is aligned at the document and sentence level. It contains about 2700 parallel documents totalling over 150,000 aligned sentences per language [9].

2.2 Multilingual parallel corpora

Concerning multilingual parallel corpora, **OPUS** is probably the largest collection of freely available parallel corpora in different languages with a considerable size and variety [10]. For example, it contains the **EuroParl** [11] and the **JRC-Acquis** [12] corpora. These two corpora both contain the EU documents of mostly legal nature such as the proceedings of the European Parliament. Both corpora are also available with bilingual alignments in all language pairs, including English. However, the EuroParl exists only in eleven European languages and contains none of the languages of the other member states or of a candidate country. Altogether, this corpus contains about 30 million words for each of the 11 languages. On the other hand, the JRC-Acquis is available in 21 official EU languages with an average size of roughly 9 million words per language. A similar corpus called **MultiUN** (A Multilingual Corpus from United Nation Documents) [13] has been extracted from the official documents of the United Nations (UN). In addition to the German language, this corpus is available in all six official languages of the UN, comprising around 300 million words per language except for German that comprises about 9 million words [14].

The following corpora are much more related to the MulTed:

- **SwissAdmin** [15]: It is one of the few freely available multilingual and PoS tagged parallel corpora. It is built of articles released by the Swiss Federal Administration. The corpus is available in four languages (English, German, Italian and French). It is released in three versions: plain texts of approximately six to eight million words per language, sentence-aligned bilingual texts for each language pair, and a PoS tagged version. The annotation has been performed automatically by the ***Fips*** multilingual parser [16].

- **AMARA** corpus [17]: It is a parallel corpus of educational video subtitles, multilingually aligned for 20 languages, i.e., 20 monolingual corpora and 190 parallel corpora. The data of this corpus were collected in cooperation with the Amara platform* using an in-house crawler. 3000 videos have subtitles available in at least six languages and 1000 videos have subtitles available in 25 languages. However, the corpus is not PoS tagged and is not freely available.

- **"WIT^3"** project [18], an acronym for **W**eb **I**nventory of **T**ranscribed and **T**ranslated **T**alks, is a collection of lecture translations that have been automatically crawled from the TED talks in a variety of languages. The purpose of this project is to support the machine translation evaluations campaigns of the International Workshop on Spoken Language Translation (IWSLT) [19]. As of October 2011, 17 thousand transcripts corresponding to translations of around 1000 talks have been crawled using the HLT Web Manager.† The latest version of "WIT^3", April 2016, contains over 2000 talks and 109 language.‡ The corpus files are classified according to the language of translation where all translations of each language are merged in one XML file. The XML files contain tags to provide information about the talk, such as its id, title, speaker, and the time slot of each segment. However, the corpus is likely to be a haphazard is a collection

of subtitles that are not balanced or classified based on the variety of TED topics. The names of the translators and reviewers are missing in the XML files. Besides, the subtitles are neither sentence-aligned nor PoS tagged.

Based on what has been described above, we propose a new corpus to address the limits of the mentioned corpora. i.e., a freely available multilingual parallel corpus, sentence-aligned, PoS tagged, covering under-resourced languages from different families, and well balanced in terms of domains and topics.

3. Data collection procedure

In this section, we present the data resource and the tools used for the collection and filtering processes. Typically, Internet users provide subtitles in various languages voluntarily. Thus, huge online databases for subtitles are available for free on the web. Sometimes, translators provide different subtitle versions of the same language for the same videos. What's more, subtitles are different from other parallel resources in various aspects, since most of them are transcriptions of spontaneous speech. Thus, they can easily be linked to the actual sound signals [1]. One of the most relevant and available subtitles on the web are those provided for TED talks.

3.1 TED talks

"TED talks" is a library of talks, filmed at independently non-profit organized events in over 130 countries. Due to the popularity of TED events worldwide, presenting high-quality content on many different topics, amazing efforts have been undertaken by at least 25,000 volunteers to generate about 40,000 translations into 101 languages or more [17]. The TED website[§] makes the video recording of the best talks and all their subtitles available under the Creative Commons BYNC-ND license. The talks are presented in an excellent and original style by very skilled speakers who cover a wide variety of topics under the slogan of "Ideas worth spreading". The talks are divided according to the languages, topics, countries and posted dates. As for the translation process, using TED talks imply dealing with spoken language, which is structurally less complex, formal and fluent than written language. Furthermore, the translators are required to follow the structure and rhythm (i.e. timing) of the English version, as it is explained in the TED platform,[**] to avoid the usual rephrasing and reordering tasks in the ordinary translation of written documents. For instance, a subtitle must not contain the end of one sentence and the beginning of another, it should be synchronized with the talk, unless the duration of a subtitle must be extended for a good reading speed.

One of the reasons that we use TED subtitles is the high-quality of their translation. As reported on its platform,[††] the subtitles go through the following steps before publication:

- **Transcription:** TED provides an original transcript.
- **Translation:** Subtitles are translated from the original language into the target language, using a simple online interface.
- **Review:** Subtitles are reviewed by an experienced volunteer (someone who has subtitled 90 min of talk content).
- **Approval:** Before publication, reviewed translations are approved by a TED Language Coordinator or staff member.

3.2 Data collection tools

For many languages, the small number of volunteers cannot keep up with the fast pace in which recent content is appearing on the TED website. Thus, we did not collect all the list of

available videos. The crawling yielded over 30,000 translations, corresponding to 1100 videos in 101 different languages. The initial collection was completed between May 10 and June 20th, 2016.

For the data collection, Google2SRT[‡‡] is used to retrieve subtitles automatically in SRT file (.srt) format. Google2SRT is a freely available tool which allows downloading, saving and converting multiple subtitles and translations from YouTube and Google Video to SubRip format. Google2SRT can extract subtitles from XML files as well as from a direct video's hyperlink or a list of video URLs saved in a TXT file. One of its useful features is the ability to select the translations that include multiple versions of the same language. It also allows choosing and saving the preferred languages in one folder in addition to the original transcript. The crawler HLTWebManager could also be used for subtitles extracting [18]; however, its use leads to an additional process to collect and link the original transcript to its translations.

3.3 Filtering and topic classification

TED talks cover a wide range of domains and topics, but not all videos come with a considerable number of translations. Thus, only the resources that met our criteria have been selected. To do so, we manually:

- Selected talks with subtitles of more than specific 15 languages, especially poor or medium density languages. We also made sure to select languages from different families such as the six official languages of the United Nations which are Arabic, Chinese, English, French, Russian and Spanish.
- Selected and organized the talks from a variety of topics to insure heterogeneity and equilibrium in the corpus. We choose the 11 following topics: "Architecture and Design", "Art and Creativity", "Culture and Stories", "Economic and Innovation", "Education and Learning", "Global Issues", "Health and Medicine", "Nature and Environment", "Science and Tech", "Social Issues" and "Sports and Adventure". From each topic, 100 videos were selected.

4. Sentence-alignment

The aim of this section is to present the sentence-alignment method used in this work. Typically, when the data are harvested, they are probably noisy. i.e., subtitles could contain wrong or incomplete components. Cleaning the noisy parallel data by detecting and removing incorrect alignments can improve the performances [20]. As a result, the content of collected subtitles has been reviewed carefully using attributes in the SRT file which include the talk and sentence IDs as well as the time-slot which is the start and end times of the segment.

Since all subtitles are segmented based on sound, there is one "segment ID" for every caption that appears on the screen at a specific timeframe. Consequently, the first method is based on time slots segmentation. Thus, the content of subtitles is segment-aligned using the segment "IDs". An evaluation of the sentence-alignment process is done implicitly; in fact, we implemented an automatic verification that discards subtitles where segment-based alignment process failed, which is the case of 2% of the subtitles. This is caused when the sequences of segments "IDs" or the total number of segments differ from those of the English transcriptions. Moreover, a manual evaluation, that involves 10 different talks from each category, was performed on two language pair, i.e. English-Arabic and English-French and no mis-aligned sentences were found.

Note that a segment could be either a whole sentence or a part of it. Considering that the subtitles contain proper punctuation, a second alignment method is applied using the strong

punctuation marks as boundaries (i.e., points and Question marks) to form a complete sentence. To do so, the sentences are regenerated by concatenating on both sides' consecutive segments until a strong punctuation mark is detected on the target side. Finally, the entire corpus is sentence-aligned considering English as a pivot language and the average number of sentences obtained is about 1.5 million per language. Figure 1 presents a sample for more illustration. It is worth mentioning that other recent works could be used for sentence-alignment like [21,22] especially texts with different complexity levels.

5. PoS tagging

In the last decades, probabilistic methods came into existence and gained more popularity because they require much less human effort. To our knowledge, the most relevant probabilistic methods used for PoS tagging are Hidden Markov Models (HMM), Support Vector Machines (SVM), and Decision Tree (DT). These methods are data-driven which means that they learn from pre-annotated corpora. From these corpora, they extract probabilities where the training task consists of learning lexical probabilities and contextual probabilities. The finest freely available language-independent PoS taggers based on these methods and with a considerable accuracy are TnT [23], SVMTool [24], and Treetagger [5].

For our purposes, we selected the Treetagger, which is probably the most widely used language-independent PoS tagger. As it can annotate texts in about 30 languages, the PoS tagging process involves only those languages for which parameter files are available in the Treetagger website,[§§] unlike the sentence-alignment procedure that includes all the 101 languages covered in the MulTed corpus.

The reported accuracy of Treetagger was about 95% (e.g., English 96.36% [5], German 97.53% [25], Russian 97.31% [26], Classical Latin 95.5% [27], and Arabic 94.7% [28]).

```
<?xml version="1.0" encoding="UTF-8"?>
<Talk id="Fg_JcKSHUtQ">
  <Category>Architecture and Design</Category >
  <Title>A robot that flies like a bird</Title>
  <Speaker>Markus Fischer</Speaker>
  <Time-slot>00:00:15,259 --> 00:00:18,259</Time-slot>
    <Segment id="1">
      <Original_text Lang_id="en">It is a dream of mankind,
      </Original_text>
        <Translation Lang_id="fr" Translator="Elisabeth Buffard"
        Reviewer="Alban Lefebvre">C'est un rêve de l'humanité,
        </Translation>
        <Translation Lang_id="es" Translator="Veronica Martinez
        Starnes" Reviewer="Sebastian Betti">Es un sueño de la
        humanidad, </Translation>
        <Translation Lang_id="ar" Translator="Faisal Jeber"
        Reviewer="Anwar Dafa-Alla">هو حلم البشرية، </Translation>
        <Translation Lang_id="zh-TW" Translator="Chunxiang
        Qian" Reviewer="Angelia King">是人类的一个梦想
        </Translation>
        <Translation Lang_id="th" Translator="Heartfelt Grace"
        Reviewer="Paravee Asava-Anan"> เป็นความใฝ่ฝันของมนุษย์,
        </Translation>
      …
    </Segment>
  …
</Talk>
```

Figure 1.
A sample of a sentence-aligned subtitle in XML format.

A sample of 500,000 words of the Arabic part of the MulTed corpus is used to evaluate the performance of Treetagger as well as TnT and SVMTool taggers. The reported accuracy rate is 88.87% as observed in Table 1.

Since the Treetagger is separately adapted to different languages, the used tagsets for each language are not identical. To annotate the MulTed corpus with a common set of tags the universal tagset [6] was adopted. Since EAGLES recommendations for the morphosyntactic annotation of corpora report that there are at least 12 main categories[***] considered obligatory for most languages, the mapping of the tagsets used for each language was done manually to convert the subcategories to the main categories. Table 2 exhibits these tags.

6. Statistical information

The current version of the MulTed corpus has subtitles covering approximately 1100 talks available in over 100 languages. The corpus comprises 30,000+ subtitles that contain 7.6 million aligned sentences with altogether over 46 million tokens. We sentence-aligned the entire corpus considering English as a pivot language, i.e., the alignment is done between English and the other languages. Then, the subtitles were classified manually into 11 categories based on the variety of TED topics. In fact, these 11 categories are a summary of those used in TED website. As noticed in the TED website,[†††] there are about 450 topics, and each talk can be associated to one or more topic. For this reason, we decided to reduce the number of categories and, accordingly, to re-classify the talks manually. Finally, after collecting, filtering, sentence-aligning, and tagging the data, we were left with the following database presented in Table 3.

Table 1. Taggers accuracies on the Arabic part of MulTed corpus.

Taggers	TnT	Treetagger	SVMTool
Accuracy	88.88%	88.87%	88.39%

Table 2. The basic tags of the universal tagset.

Tags	Tag symbols
Verbs	VERB
Nouns	NOUN
Pronouns	PRON
Adjectives	ADJ
Adverbs	ADV
Determiners (determiners and articles)	DET
Adpositions (prepositions and postpositions)	ADP
Particles	PRT
Numerals	NUM
Conjunctions	CON
Other (typos, abbreviations. . .)	X
Punctuation marks	SENT

Table 3. Detailed information about MulTed corpus.

	Nb. of talks	Nb. of languages	Nb. of subtitles	Nb. of segments	Nb. of tokens
Total	1100	101	30,057	7.6 million	46 + million

To estimate the number of segments and tokens, the following pre-processing tasks were performed:

- **Normalization:** to remove non-needed subtitle components (i.e. time slots and talk id. . .) and special characters except punctuation marks;
- **Tokenization:** to break up the text into individual tokens using as delimiters, white-space, and newline.

In term of languages, there are languages that have subtitles in a considerable number of videos such as English (en) with 1100 subtitles, Arabic (ar) with 1090 subtitles, Brazilian Portuguese (pt-BR) and Hebrew (iw) with 1085 subtitles for each language, Korean (ko) with 1080 subtitles, and French (fr) with 1074 subtitles. Besides, many resource-poor languages are covered in the corpus such as Thai (th) with 624 subtitles and Indonesian (id) with 509 subtitles. Figure 2 presents the overall distribution of all 101 languages grouped by sizes of the number of talks.

Next, we present more details about these top 30 languages. Table 4 displays the number of monolingual files, the number of tokens per language, and the number of segments pairs with English, respectively. Note that these languages are relatively different from the PoS tagged languages discussed in Section 5. For instance, languages like Hebrew, Farsi, Indonesian, and Japanese are among the top 30 languages, but they did not receive PoS tags since their parameter files are not available yet on the Treetagger website.

7. Corpus format

In addition to the original format, XML, as an encoding language, is used to facilitate the use of the corpus. Thus, this corpus is released in three versions:

1. Plain text version of approximately 0.6 (e.g., Slovenian) to 2.1 (e.g., English) million tokens per language;
2. Sentence-aligned bilingual texts version for each language pair;
3. PoS tagged version for each talk.

Moreover, the XML files in all versions contain tags and attributes to provide further metadata of the talk. For instance, talk id, title, category, translator, time slot, speaker, and language id. Their meaning is self-explanatory. Figure 3 exhibits a sample of the PoS tagged version of an English subtitle.

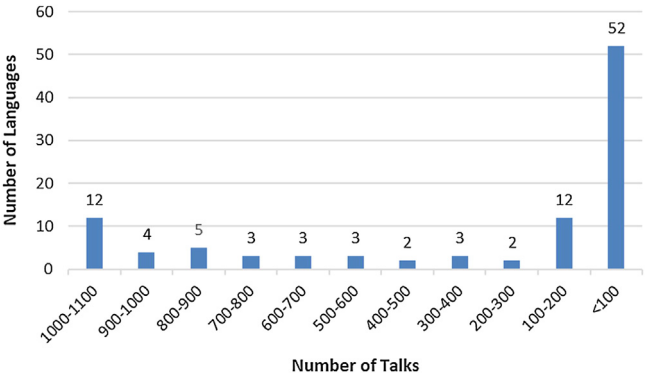

Figure 2.
Distribution of the 101 languages by the number of talks.

Languages	Nb. of files	Nb. of tokens	Nb. of Segments
English	1100	2,134,155	275,847
Arabic	1090	1,703,114	271,246
Portuguese-BR	1085	1,978,894	270,255
Hebrew	1085	1,568,768	270,534
Korean	1080	1,417,473	266,449
French	1074	2,124,021	267,901
Russian	1071	1,684,719	270,825
Spanish	1068	1,968,660	266,847
Chinese (TW)	1065	326,739	265,535
Italian	1048	1,859,765	261,477
Japanese	1045	406,964	264,610
Romanian	1003	1,750,662	242,909
Chinese (CN)	977	298,569	258,122
Dutch	957	1,726,997	240,410
Vietnamese	944	2,279,456	232,153
Polish	932	1,297,125	228,738
Greek	898	1,582,023	222,279
Deutsch	894	1,604,856	226,573
Turkish	885	1,239,183	224,495
Serbian	871	1,378,420	212,990
Hungarian	800	1,188,401	202,221
Bulgarian	791	1,425,035	205,132
Portuguese	790	1,435,525	199,333
Farsi	787	1,673,768	193,726
Ukrainian	680	1,016,555	164,927
Croatian	648	1,019,459	158,525
Thai	624	270,437	145,432
Czech	581	880,445	139,670
Indonesian	509	764,600	113,990
Slovenian	494	607,375	96,943

Table 4.
The number of segments pairs with English.

```
<?xml version="1.0" encoding="UTF-8"?>
<Talk id="Fg_JcKSHUtQ">
  <Category>Architecture and Design</Category >
  <Title>A robot that flies like a bird</Title>
  <Speaker>Markus Fischer</Speaker>
  <Time-slot>00:00:15,259 --> 00:00:18,259</Time-slot>
    <Segment id="1">
      <Word PoS="PRON" Lemma="It">It</Word>
      <Word PoS="VERB" Lemma="be">is</Word>
      <Word PoS="DET" Lemma="a">a</Word>
      <Word PoS="NOUN" Lemma="dream">dream</Word>
      <Word PoS="ADP" Lemma="of">of</Word>
      <Word PoS="NOUN" Lemma="mankind">mankind</Word>
      <Word PoS="SENT" Lemma="unknown">,</Word>
    </Segment>
    …
</Talk>
```

Figure 3.
A sample of a PoS tagged version of an English subtitle.

8. Discussion

The major drawback of many multilingual parallel corpora is probably that they are compiled based mainly on legislative or technical raw data (e.g., SwissAdmin and MultiUN). Also, these corpora are restricted mostly to high-density languages such as English and the

official European languages (e.g., EuroParl). Besides, some of them are in fact bilingual pairs rather than multilingual (e.g., CzEng). Regarding those corpora that consist of subtitles, most of them target movies and TV shows subtitles. This creates a challenge for sentence-alignment due to the possibility of multi-speakers in the same time slot. Further, the language used in movies is informal. Thus, translating this language is much more challenging since research in SMT has mostly been driven by formal translation tasks [29]. Another investigation reports that movies subtitles are not well-suited for machine translation purpose [30]. In addition, several parallel corpora are neither balanced in terms of topics covered, nor sentence-aligned or PoS tagged. As for the MulTed corpus, it is not aligned on a word level and is not PoS tagged for 71 languages, such as Japanese and Indonesian.

Nevertheless, Tiedemann et al. [31] confirms that the quality of the training data is essential to increase translation performance. Therefore, the MulTed corpus has some significant advantages such as the high-quality of its translations. In fact, the crowdsourced transcriptions and the translations are reviewed by experienced translators and then approved by TED Language Coordinators and staff members. Additionally, the TED talks are presented in well-structured language which makes this kind of corpus very valuable to build SMT systems. Indeed, since 2011, the transcriptions and the translations of TED talks are used yearly, as training and testing data, for an open evaluation campaign on spoken language translation in the International Workshop on Spoken Language Translation (IWSLT[‡‡‡]). Furthermore, the MulTed corpus covers a variety of topics in the used raw database. Unlike "WIT[3]", the subtitles of the MulTed corpus are classified and balanced manually into 11 categories based on these topics. Moreover, this corpus is characterized by language diversity since it covers high, medium, and poor density languages from different families. Besides, the MulTed is based on talks that have only one speaker which helps the alignment process as well as text compression and summarization studies, as done in the European projects MUSA and Flemish ATraNoS to summarize the discussion of TV shows [32]. Finally, it is a multilingual parallel corpus that covers over 100 languages, sentence-aligned, and PoS tagged.

9. Conclusion and future works

This paper has shed light on the characteristics of the MulTed corpus, a new multilingual and PoS tagged parallel corpus with bilingual sentence alignment considering English as a pivot language. The corpus has been derived from the TED talks, where volunteers contribute transcriptions and translations that are available to the public. The corpus currently contains subtitles that cover 1100 talks, including a variety of domains and topics; yet, it is characterized by language diversity where at least 30 languages are well covered. In addition to the sentence alignment, the datasets are annotated based on a language independent part of speech tagger and a universal common tagset. Also, the data has been stored in a uniform XML format. Hence, as a research purpose, the MulTed corpus will be released freely under a Creative Commons licence via our team website[§§§] with future updates as new talks and translations are released. This will make the corpus useful for applications in computational linguistics, cross-linguistic studies, and statistical machine translation systems, etc.

In the future, the authors are particularly interested in extending the corpus to cover more languages, new transcripts, and translations from other educational references (e.g., Khan Academy and Coursera). The next releases of the corpus will include more updates as recent translations become available. The translation can bridge the language gap by enabling users to benefit from a valuable content on the web. Therefore, we plan to explore the usage of MulTed corpus in machine translation which can facilitate the task of the translators by performing post-editing instead of starting from scratch. There's more, the files of MulTed corpus are classified based on their topics. Consequently, this corpus needs to be considered

for use in both training and testing Topic Detection and Tracking methods that aim to locate topically related documents in streams of data. Finally, since the subtitles are transcripts of a speech, they can easily be linked to the actual sound signals. Thus, including the sound signals and link them to their segments is applicable which will make the corpus more valuable for speech synthesis and speech recognition.

Notes

* http://amara.org.

† https://wit3.fbk.eu/tools/WebManager_Manual.pdf.

‡ https://wit3.fbk.eu/#lastestver.

§ http://TEDtalks.ted.com.

** http://translations.ted.org/wiki.

†† https://www.ted.com/participate/translate/get-started.

‡‡ https://sourceforge.net/projects/google2srt/.

§§ http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

*** http://www.ilc.cnr.it/EAGLES96/annotate/node16.html#cmobli.

††† https://www.ted.com/topics.

‡‡‡ http://iwslt.org.

§§§ http://oujda-nlp-team.net/en/corpora/multed-corpus/.

References

[1] J. Tiedemann, Building a multilingual parallel subtitle corpus, in: Proceedings of CLIN 17, Leuven, Belgium, 2007.

[2] S. Green, J. Heer, C.D. Manning, The efficacy of human post-editing for language translation, in: Proc. SIGCHI Conf. Hum. Factors Comput. Syst., ACM, 2013: pp. 439–448.

[3] T. Naseem, B. Snyder, J. Eisenstein, R. Barzilay, Multilingual part-of-speech tagging: two unsupervised approaches, J. Artif. Intell. Res. 36 (2009) 341–385.

[4] J. Sylak-Glassman, C. Kirov, M. Post, R. Que, D. Yarowsky, A universal feature schema for rich morphological annotation and fine-grained cross-lingual part-of-speech tagging, in: Int. Workshop Syst. Framew. Comput. Morphol., Springer, 2015, pp. 72–93.

[5] H. Schmid, Probabilistic part-ofispeech tagging using decision trees, in: New Methods Lang. Process., Routledge, 2013, p. 154.

[6] S. Petrov, D. Das, R. McDonald, A universal part-of-speech tagset, in: Proc. Eighth Int. Conf. Lang. Resour. Eval. LREC-2012, 2012.

[7] O. Bojar, O. Dušek, T. Kocmi, J. Libovickỳ, M. Novák, M. Popel, R. Sudarikov, D. Variš, Czeng 1.6: enlarged czech-english parallel corpus with processing tools dockered, in: Int. Conf. Text Speech Dialogue, Springer, 2016, pp. 231–238.

[8] M. Neves, A.J. Yepes, A. Névéol, The scielo corpus: a parallel corpus of scientific publications for biomedicine, Proc. Tenth Int. Conf. Lang. Resour. Eval. LREC 2016 Paris Fr. Eur. Lang. Resour. Assoc., ELRA, 2016.

[9] W. Aziz, L. Specia, Fully automatic compilation of Portuguese-English and Portuguese-Spanish Parallel Corpora, in: Proc. 8th Braz. Symp. Inf. Hum. Lang. Technol. STIL 2011 Cuiabá, 2011, pp. 234–238.

[10] J. Tiedemann, Parallel Data, Tools and Interfaces in OPUS, in: LREC, 2012, pp. 2214–2218.

[11] P. Koehn, Europarl: A parallel corpus for statistical machine translation, in: MT Summit, Citeseer, 2005, pp. 79–86.

[12] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, D. Varga, The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages, in: Proc. Fifth Int. Conf. Lang. Resour. Eval. LREC 2006, European Language Resources Association (ELRA), Genoa, Italy, 2006, p. 2006.

[13] Y. Chen, A. Eisele, MultiUN v2: UN documents with multilingual alignments, in: LREC, 2012, pp. 2500–2504.

[14] R. Steinberger, M. Ebrahim, A. Poulis, M. Carrasco-Benitez, P. Schlüter, M. Przybyszewski, S. Gilbro, An overview of the European Union's highly multilingual parallel corpora, Lang. Resour. Eval. 48 (2014) 679–707.

[15] Y. Scherrer, L. Nerima, L. Russo, M. Ivanova, E. Wehrli, SwissAdmin: A multilingual tagged parallel corpus of press releases, in: Proc. Ninth Int. Conf. Lang. Resour. Eval. LREC-2014, 2014.

[16] E. Wehrli, L. Nerima, The Fips multilingual parser, in: Lang. Prod. Cogn. Lex., Springer, 2015, pp. 473–490.

[17] A. Abdelali, F. Guzman, H. Sajjad, S. Vogel, The AMARA corpus: building parallel language resources for the educational domain, in: LREC, 2014, pp. 1044–1054.

[18] M. Cettolo, C. Girardi, M. Federico, Wit3: Web inventory of transcribed and translated talks, in: Proc. 16th Conf. Eur. Assoc. Mach. Transl. EAMT, 2012, pp. 261–268.

[19] M. Paul, M. Federico, S. Stüker, Overview of the IWSLT 2010 evaluation campaign, in: IWSLT, 2010, pp. 3–27.

[20] C. Goutte, M. Carpuat, G. Foster, The impact of sentence alignment errors on phrase-based machine translation performance, in: Proc AMTA, 2012.

[21] S. Štajner, M. Franco-Salvador, P. Rosso, S.P. Ponzetto, CATS: a tool for customized alignment of text simplification corpora, in: N.C. Conference chair, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proc. Elev. Int. Conf. Lang. Resour. Eval. LREC 2018, European Language Resources Association (ELRA), Miyazaki, Japan, 2018.

[22] S. Štajner, M. Franco-Salvador, S.P. Ponzetto, P. Rosso, H. Stuckenschmidt, Sentence alignment methods for improving text simplification systems, in: Proc. 55th Annu. Meet. Assoc. Comput. Linguist. Vol. 2 Short Pap., 2017, pp. 97–102.

[23] T. Brants, TnT, a statistical part-of-speech tagger, Proc. Sixth Conf. Appl. Nat. Lang. Process., Association for Computational Linguistics, 2000, pp. 224–231.

[24] L. Màrquez, J. Giménez, A general pos tagger generator based on support vector machines, J. Mach. Learn. Res. (2004).

[25] H. Schmid, Improvements in part-of-speech tagging with an application to German, in: Nat. Lang. Process. Using Very Large Corpora, Springer, 1999, pp. 13–25.

[26] E. Kotelnikov, E. Razova, I. Fishcheva, A close look at Russian morphological parsers: which one is the best?, in: Conf Artif. Intell. Nat. Lang., Springer, 2017, pp. 131–142.

[27] A. Field, An automated approach to syntax-based analysis of classical Latin, Digit. Class. Online 3 (2016) 57–78.

[28] I. Zeroual, A. Lakhouaja, R. Belahbib, Towards a standard Part of Speech tagset for the Arabic language, J. King Saud Univ. Comput. Inf. Sci. 29 (2017) 174–181.

[29] M. van der Wees, A. Bisazza, C. Monz, Measuring the Effect of Conversational Aspects on Machine Translation Quality, COLING, 2016.

[30] A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing: 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012, Springer Science & Business Media, 2012, vol. 7181.

[31] J. Tiedemann, F. Cap, J. Kanerva, F. Ginter, S. Stymne, R. Ostling, M. Weller-Di Marco, Phrase-Based SMT for Finnish with More Data, Better Models and Alternative Alignment and Translation Tools, in: Proc. First Conf. Mach. Transl. Berl. Ger. Assoc. Comput. Linguist., 2016.

[32] W. Daelemans, A. Höthker, E.F.T.K. Sang, Automatic Sentence Simplification for Subtitling in Dutch and English., in: LREC, 2004.

Corresponding author
Imad Zeroual can be contacted at: mr.imadine@gmail.com