WIT^3: Web Inventory of Transcribed and Translated Talks

Mauro Cettolo **Christian Girardi** **Marcello Federico**

FBK – Fondazione Bruno Kessler
Trento, Italy
{surname}@fbk.eu

Abstract

We describe here a Web inventory named WIT^3 that offers access to a collection of transcribed and translated talks. The core of WIT^3 is the TED Talks corpus, that basically redistributes the original content published by the TED Conference website (http://www.ted.com). Since 2007, the TED Conference, based in California, has been posting all video recordings of its talks together with subtitles in English and their translations in more than 80 languages. Aside from its cultural and social relevance, this content, which is published under the Creative Commons BY-NC-ND license, also represents a precious language resource for the machine translation research community, thanks to its size, variety of topics, and covered languages. This effort repurposes the original content in a way which is more convenient for machine translation researchers.

1 Introduction

Data play a key role in machine learning as they are the main source of information to infer parameter values of the employed mathematical model.

In statistical machine translation (SMT), learning is performed on parallel texts, i.e. documents, sentences or even fragments of sentences with their translation(s). Large amounts of in-domain parallel data are usually required to properly train translation and reordering models.

Unfortunately, parallel data are a scarce resource, which are freely available only for some language pairs and for few, very specific domains. For example, MultiUN (Eisele and Chen, 2010) provides large parallel texts (300 million words) but for only 6 languages; Europarl (Koehn, 2005) consists of the translation into most European languages of the proceedings of the European Parliament (at most 50 million words); JRC-Acquis[1] comprises the total body of European Union law applicable to the Member States, written in 22 European languages (35 million words); other smaller parallel corpora in specific domains are included in OPUS (Tiedemann, 2009) for various languages.

On the other hand, it is unfeasible for research laboratories to cover all possible needs in terms of parallel texts by resorting to professional translators, given their high cost.

The data available at the TED website[2] is therefore particularly valuable for the MT community. TED is a nonprofit organization that invites "the world's most fascinating thinkers and doers [...] to give the talk of their lives". The site makes available under the Creative Commons BY-NC-ND license the video recordings of the best TED talks, all subtitled in English and translated in many other languages by volunteers worldwide. The set of subtitles represents a precious multilingual parallel corpus since its size continuously increases (more than 900 TED talks had been collected at the end of 2011), subtitles are available in a significant number of languages (82 now, to be extended to 90 in the near future) and topics covered span the whole of human knowledge, making such data useful for any possible application domain.

In order to make this collection of talks more effectively usable by the research community, we

[1] http://optima.jrc.it/Acquis (accessed April 16, 2012).
[2] http://www.ted.com (accessed April 16, 2012).

have developed WIT^3 – an acronym for Web Inventory of Transcribed and Translated Talks –, a website hosting a ready-to-use version of this multilingual corpus, benchmarks for MT based on these data, as well as software tools to process them.

The paper is organized as follows: The TED Talks corpus is presented in Section 2, where specific subsections are devoted to the format of the files and to the sentence-level alignment; corpus statistics and an objective analysis of the difficulty of translating TED talks are also given. Section 3 describes the use of the TED Talks Corpus in the MT evaluations campaigns of the International Workshop on Spoken Language Translation (IWSLT). Finally, experimental results on baseline systems developed on several language pairs are provided in Section 4. The paper ends with the description of the WIT^3 website (Section 5) and a summary (Section 6).

2 TED Talks Corpus

TED talks are mostly held in English and their videos are available through the TED website together with subtitles provided in many languages. Almost all of the talks have been translated, by volunteers, into Arabic, Bulgarian, Chinese (simplified), French, Italian, Korean, Portuguese (Brazil) and Spanish. For about 70 other languages, the number of translated talks ranges from several hundreds (e.g. such as other Dutch, German, Hebrew, Romanian) to one (e.g. Hausa, Hupa, Bislama, Ingush, Maltese). Notice that original subtitles and their translations are segmented on the basis of sound, hence the correspondence between captions and sentences is weak. It may therefore happen both that sentences are split into more consecutive captions, and that captions include sentences fragments.

For preparing parallel corpora, the raw data were first crawled, translations of the same talks were paired, captions were aligned and sentences were re-built. Each single step is described in some detail in the following subsections.

2.1 Crawling

TED talk subtitles are crawled by means of HLTWebManager (Girardi, 2011), an in-house crawler written in Java for downloading pages published on the Web in different languages. From the original HTML downloaded documents, only subtitles and useful metadata concerning talks are kept and stored in a XML format defined by the DTD available at the WIT^3 website (Section 5). For each language, a single XML file is generated which includes all talks subtitled in that language. Each talk is enclosed in tags `<file id="int">` and `</file>` and includes, among other tags:

`<url>`	the address of the original HTML document of the talk
`<speaker>`	the name of the talk speaker
`<talkid>`	the numeric talk identifier
`<transcript>`	talk subtitles split in captions
`<date>`	the issue date of the talk
`<content>`	talk subtitles

The `transcript` and `content` fields only differ in the presence of timestamps indicating splits introduced to make subtitles readable during video playing.

The `talkid` field is an integer uniquely identifying the original transcript of a talk and all its translations. Therefore, it can be used to pair translations of the same talk.

There are other tags (e.g. `description`, `keywords`, `title`, whose meaning is self-explanatory) that, providing further metadata of the talks, could be exploited for purposes like clustering, information retrieval, categorization and adaptation.

2.2 Alignment

Given a pair of languages, it is straightforward to select the talks for which subtitles are available in both languages, exploiting the `talkid` mentioned in Section 2.1. For each of such talks, the captions in the two languages are extracted from the `transcript` tags and paired in the order of appearance. A number of heuristic checks are performed in order to assess the parallelism. A whole talk is discarded if either the number of captions in the two documents differs, or the sequences of timestamps differ. Moreover, pairs of aligned captions within a talk are marked as unreliable and removed if their length ratio is an outlier, assuming a normal distribution and a 95% confidence interval.

To get an idea of the impact of filtering data with these heuristics, for the English–French collection it eliminates about 3% of the words.

Once captions are aligned, sentences are regenerated by concatenating on both sides consecutive captions until a strong punctuation mark is

	ar	bg	zh	en	fr	it	ko	pt-BR	es	de	he	nl	pl	ro	ru	tr	cs	el	hu	ja	fa	pt	vi
ar	-	1.29	1.28	1.36	1.31	1.19	1.21	1.31	1.34	0.79	0.95	0.65	0.86	1.02	0.75	0.95	0.41	0.48	0.63	0.69	0.34	0.41	0.45
bg	1.39	-	1.63	1.72	1.61	1.47	1.46	1.67	1.70	0.91	1.13	0.75	1.02	1.22	0.86	1.12	0.49	0.55	0.71	0.81	0.37	0.46	0.51
zh	0.21	0.24	-	0.25	0.24	0.21	0.22	0.24	0.25	0.13	0.17	0.11	0.15	0.18	0.13	0.17	0.07	0.08	0.11	0.12	0.06	0.07	0.08
en	1.63	1.91	1.89	-	1.92	1.70	1.70	1.93	1.98	1.08	1.35	0.92	1.20	1.43	1.01	1.35	0.59	0.65	0.85	0.94	0.48	0.56	0.62
fr	1.64	1.87	1.86	2.00	-	1.69	1.70	1.91	1.96	1.06	1.31	0.87	1.17	1.40	0.99	1.31	0.55	0.64	0.83	0.93	0.46	0.54	0.60
it	1.34	1.53	1.48	1.60	1.52	-	1.37	1.54	1.57	0.93	1.13	0.78	1.00	1.23	0.90	1.11	0.50	0.57	0.74	0.82	0.41	0.48	0.52
ko	1.01	1.12	1.11	1.17	1.12	1.01	-	1.14	1.14	0.66	0.80	0.53	0.73	0.87	0.63	0.81	0.33	0.40	0.51	0.59	0.28	0.34	0.37
pt-BR	1.51	1.78	1.76	1.85	1.76	1.58	1.60	-	1.84	0.98	1.23	0.80	1.06	1.31	0.93	1.20	0.52	0.60	0.76	0.87	0.41	0.50	0.55
es	1.56	1.84	1.80	1.92	1.82	1.63	1.62	1.86	-	1.00	1.26	0.83	1.09	1.35	0.95	1.24	0.53	0.60	0.79	0.88	0.42	0.50	0.57
de	0.91	0.96	0.94	1.03	0.97	0.95	0.91	0.97	0.98	-	0.84	0.62	0.75	0.87	0.71	0.79	0.41	0.47	0.58	0.63	0.35	0.39	0.42
he	0.85	0.93	0.91	1.00	0.93	0.89	0.86	0.95	0.96	0.65	-	0.54	0.67	0.80	0.63	0.73	0.36	0.41	0.49	0.55	0.29	0.34	0.36
nl	0.75	0.81	0.79	0.89	0.81	0.81	0.74	0.81	0.83	0.63	0.71	-	0.64	0.74	0.58	0.69	0.40	0.40	0.51	0.50	0.32	0.35	0.39
pl	0.78	0.86	0.84	0.90	0.85	0.80	0.80	0.83	0.84	0.59	0.68	0.50	-	0.70	0.55	0.67	0.30	0.35	0.47	0.51	0.26	0.30	0.34
ro	1.19	1.31	1.28	1.38	1.30	1.27	1.22	1.32	1.35	0.88	1.05	0.74	0.90	-	0.82	1.04	0.50	0.55	0.69	0.77	0.40	0.45	0.51
ru	0.75	0.80	0.80	0.85	0.80	0.80	0.77	0.81	0.82	0.62	0.71	0.50	0.61	0.71	-	0.67	0.35	0.39	0.49	0.52	0.30	0.31	0.36
tr	0.82	0.89	0.88	0.97	0.90	0.85	0.85	0.90	0.92	0.60	0.71	0.51	0.65	0.77	0.57	-	0.32	0.38	0.48	0.52	0.27	0.31	0.37
cs	0.40	0.45	0.43	0.49	0.44	0.43	0.40	0.45	0.45	0.35	0.40	0.33	0.33	0.42	0.35	0.36	-	0.25	0.32	0.29	0.21	0.22	0.23
el	0.55	0.58	0.55	0.62	0.59	0.58	0.54	0.59	0.59	0.46	0.52	0.40	0.44	0.54	0.44	0.49	0.29	-	0.39	0.38	0.26	0.27	0.28
hu	0.59	0.61	0.60	0.66	0.62	0.61	0.57	0.62	0.64	0.47	0.51	0.41	0.48	0.56	0.45	0.51	0.30	0.32	-	0.41	0.24	0.26	0.27
ja	0.13	0.14	0.14	0.15	0.14	0.14	0.13	0.14	0.14	0.10	0.12	0.08	0.11	0.12	0.10	0.11	0.05	0.06	0.08	-	0.05	0.06	0.06
fa	0.47	0.47	0.50	0.57	0.51	0.50	0.47	0.49	0.50	0.43	0.46	0.39	0.38	0.47	0.41	0.44	0.30	0.32	0.36	0.38	-	0.27	0.24
pt	0.48	0.50	0.49	0.54	0.50	0.49	0.48	0.49	0.50	0.40	0.44	0.35	0.38	0.45	0.36	0.41	0.26	0.27	0.33	0.34	0.22	-	0.24
vi	0.69	0.71	0.71	0.80	0.73	0.71	0.68	0.72	0.75	0.55	0.61	0.51	0.57	0.67	0.55	0.65	0.35	0.38	0.45	0.50	0.27	0.33	-

Table 1: The names of languages are represented by ISO 639-1 codes. Numbers refer to millions of units (untokenized words). (`row,col`) entries of bottom-left triangle provide the size of parallel text available for the `row` language side, those of upper-right triangle, for the `col` language side.

detected on the target side. This means that the provided parallel corpus could have: (i) lines including more sentences, as sentences can end inside captions; (ii) source lines that do not end with a strong punctuation mark.

2.3 Statistics

As of October 2011, we have collected almost 17 thousand transcripts, corresponding to translations of around 1000 English talks into 80 languages. Crawled text in all languages is left in its original format. In particular, no tokenization is applied and no word segmentation is performed for languages such as Chinese and Japanese. Hence, the reported size of corpora refer to the number of tokens, or string units, where words are possibly joined to punctuation marks and not segmented.

The distribution of translations over the 80 languages is very uneven, and consequently even more sparse among the possible 3160 language pairs.

For the three pairs from {English, French, Spanish}, parallel data reach about 2 million units. At least 1 million units can be collected for all pairs from a set of 9 languages (36 possible pairs), while at least 500K for any pair from a set of 16 languages (120 possible pairs) and at least 200 thousand for any pair from a set of 23 languages (253 possible pairs). Table 1 collects the size of parallel corpus available for each pair from the 9/16/23 sublists.

2.4 Insights

How difficult is to translate TED talks? One hint comes from the scores obtained by participants at the recent evaluation campaign of IWSLT 2011 (Federico et al., 2011), which organized MT tracks based on the TED Talks data. The best reported automatic scores, computed on a single reference (see Table 7), are in fact comparable to those obtained by the best systems in the 2011 WMT evaluation (Callison-Burch et al., 2011) for the English–to–French direction on the generic news domain. This comparison is particularly significant given the similarity of experimental conditions: equivalent amount of in-domain training data and same out-of-domain training corpora. On the other hand, IWSLT scores for the translation of TED Talks from Arabic and Chinese into English are definitely lower than those obtained on news by the best systems in the last NIST evaluation,[3] for the same translation directions; however, in this case the comparison is made difficult by the very different training conditions and by the use of multiple references in score computation.

[3] `http://www.itl.nist.gov/iad/mig/tests/mt/2009/ResultsRelease/progress.html` (accessed April 16, 2012).

Beyond using MT performance scores, the difficulty of a translation task can be weakly related to the target language model perplexity (PP) and out-of-vocabulary word rate (OOV). If such figures are computed on in-domain data, they provide hints on how intrinsically hard the task is; if they are computed on out-of-domain texts, they provide a cue on how close and potentially useful they are to improve in-domain models.

Hence, as a case study, we analyzed the English–to–French translation track of the 2011 IWSLT evaluation campaign. First, 5-gram language models (LMs) have been estimated on a number of French texts made available for training purposes, namely:

- TED: the monolingual French corpus consisting of TED talks; it is the only in-domain text
- NC: the French side of the parallel English–French News Commentary corpus
- EPPS: the French side of the parallel English–French Europarl corpus
- MultiUN: the French side of the parallel English–French MultiUN corpus.

The PP/OOV of the target side of the 2011 English–to–French test set have then been computed using each LM and collected in Table 2, which reports also the number of tokens used for training the LMs.

data	corpus size	PP	%OOV
TED	2.35M	103.8	1.67
NC	3.36M	266.8	2.83
EPPS	56.2M	200.3	1.79
MultiUN	402.8M	288.2	1.21
all	464.7M	150.8	0.72

Table 2: PP and %OOV of the IWSLT 2011 test set with respect to four 5-gram LMs estimated on in- and out-of-domain different sized corpora. Values are also reported for the LM built on the union of all corpora.

The following considerations can be drawn:

- the in-domain corpus always gets the lowest PP, even if it is the smallest one; this shows that even if the topics covered by the TED talks are rather different, the common situation induces speakers to use a somehow similar language
- the TED talks are quite far from all the other genres of text considered here: news, proceedings of the European Parliament and resolutions of the General Assembly of the United Nations. It is quite unexpected that EPPS is closer to talks than news, but the difference in PP could be due to the size of the two corpora rather than their nature
- the OOV with respect to out-of-domain corpora seems to be mainly related to their size; it is worth noticing that the OOV can be more than halved if out-of-domain corpora are added to the in-domain one (see entry `all`), showing that the proper exploitation of available data can be anyway beneficial.

The figures just analyzed regard the evaluation set as a whole, but one could wonder if they hide large fluctuations across different talks. Table 3 provides some figures computed at talk-level both on the test set and on the TED training corpus; specifically: the mean μ of PP and OOV, their standard deviations σ, minimum and maximum values. Concerning the test set, the scores were computed on the LM estimated on text available for training. For talks in the training set, figures were computed using a 1-fold cross validation scheme.

		μ	σ	[min,max]
tst2011	PP	103.7	19.7	[68.9,132.0]
	%OOV	1.55	0.46	[0.91,2.37]
training	PP	130.2	49.3	[38.8,505.7]
	%OOV	1.76	1.04	[0.00,15.79]

Table 3: Mean, standard deviation and minimum and maximum values of PP and %OOV of talks in the test and training sets.

It results what follows:

- on average, the values of PP and OOV of talks selected for evaluation are lower than those of talks included in training data; likely, this is due to the presence of very hard talks in the training data or of very easy talks in the testing data
- the ([min,max]) ranges of observed PP and OOV values are rather large; this means

that talks can linguistically differ significantly among each others and consequently MT performance on them too.

3 WIT^3 for IWSLT evaluations

The International Workshop on Spoken Language Translation is a yearly event associated with an open evaluation campaign on spoken language translation. IWSLT proposes every year challenging research tasks and an open experimental infrastructure for the scientific community working on spoken and written language translation.

In 2010 edition (Paul et al., 2010), alongside the tasks on traveling domain built on the BTEC corpus (Takezawa et al., 2007), a new challenge was introduced, that is the translation of TED talks. This became the only MT task proposed to participants in edition 2011 (Federico et al., 2011) and will remain the main task in 2012 as well.

From a translation point of view, TALK is basically a subtitling translation task, in which the ideal translation unit is a single caption as defined by the original transcript.

Concerning training data, in the 2011 edition, in addition to the roughly 2-million word parallel corpora of TED talks for each considered language pair, several out-of-domain large parallel corpora have been provided, including texts from the United Nations, European Parliament, news commentaries and the Web.

From 2012, TED Talks training data for the IWSLT evaluations will be distributed through the WIT^3 website. In addition to the official tasks, the site will also release unofficial benchmarks for many other language pairs.

4 Baselines

In this section, we present results on some benchmarks that we obtained by training MT baseline systems on the available TED Talks data. The aim is to provide MT scientists with reference results that can help them in assessing their experimental outcomes. In addition to language pairs for which results were already published at IWSLT 2011, we have considered several new translation directions. The scores reported for the former will allow the assessment of the quality of our baselines with respect to state-of-the-art systems; the scores reported for the new languages can help either to understand the degree of difficulty of the task or simply to set a reference .

4.1 IWSLT 2011 MT Track Language Pairs

4.1.1 Data

Experiments were performed on data supplied by the organizers of the IWSLT 2011 evaluation campaign for the MT track,[4] who asked participants to automatically translate talks from Arabic to English, from Chinese to English and from English to French. For developing the baselines, only texts from the TED domain were employed, i.e. no additional out-of-domain resources were used. Different preprocessings were performed depending to the language: Arabic and Chinese were segmented by means of AMIRA (Diab et al., 2004) and the Stanford Chinese Segmenter (Tseng et al., 2005), respectively; the tokenizer script released together with the Europarl corpus (Koehn, 2005) was applied to other languages.

The same partitioning of the evaluation campaign in terms of parallel training data, development (dev2010, tst2010) and test (tst2011) sets has been adopted: Tables 4 and 5 report some statistics of such texts.

text	#sent.	Arabic		English	
		$\|W\|$	$\|V\|$	$\|W\|$	$\|V\|$
parallel	90.6k	1.71M	71.1k	1.74M	42.5k
dev2010	934	19.3k	4.6k	20.1k	3.4k
tst2010	1664	30.9k	6.0k	32.0k	3.9k
tst2011	1450	26.7k	5.8k	27.0k	3.7k
text	#sent.	Chinese		English	
		$\|W\|$	$\|V\|$	$\|W\|$	$\|V\|$
parallel	107.1k	1.95M	51.9k	2.07k	46.9k
dev2010	934	21.6k	3.7k	20.1k	3.4k
tst2010	1664	33.3k	4.4k	32.0k	3.9k
tst2011	1450	24.8k	3.9k	27.0k	3.7k
text	#sent.	English		French	
		$\|W\|$	$\|V\|$	$\|W\|$	$\|V\|$
parallel	107.3k	2.07M	46.6k	2.22M	58.2k
dev2010	934	20.1k	3.4k	20.3k	3.9k
tst2010	1664	32.0k	3.9k	33.8k	4.8k
tst2011	818	14.5k	2.5k	15.6k	3.0k

Table 4: Statistics on parallel data used for setting up the baselines of IWSLT 2011 language pairs. "#sent." stands for "number of sentences", $|W|$ for "running words", $|V|$ for "vocabulary size", k and M for 10^3 and 10^6, respectively. Counts refer to tokenized texts.

[4] `http://www.iwslt2011.org/doku.php?id=06_evaluation` (accessed April 16, 2012).

monolingual	#sent.	$\|W\|$	$\|V\|$
English	123.9k	2.41M	51.3k
French	111.4k	2.32M	60.3k

Table 5: Statistics on monolingual data used for training LMs of IWSLT 2011 target languages. See caption of Table 4 for the meaning of symbols.

4.1.2 Performance

The SMT baseline systems are built upon the open-source MT toolkit Moses (Koehn et al., 2007). The translation and the lexicalized reordering models were trained on parallel training data; taking into account the limited amount of training data, 4-gram LMs smoothed through the improved Kneser-Ney technique (Chen and Goodman, 1999) were estimated on monolingual texts via the IRSTLM toolkit (Federico et al., 2008). The weights of the log-linear interpolation models were optimized on the development sets `dev2010` by means of the standard MERT procedure provided within the Moses toolkit. Performance scores were computed with `MultEval`, using the implementation by (Clark et al., 2011).

Table 6 collects the %BLEU, METEOR and TER scores and their standard deviations ("case sensitive+punctuation" mode) of the baseline systems for the considered language pairs. In addition to the scores obtained for `dev2010` after the last iteration of the tuning algorithm, scores measured for the second development set (`tst2010`) and for the official test set (`tst2011`) of the evaluation campaign are reported.

	%bleu	σ	mtr	σ	ter	σ
ar-en						
dev2010	23.35	0.54	47.19	0.39	57.15	0.58
tst2010	22.10	0.44	46.09	0.35	59.38	0.50
tst2011	21.35	0.49	44.74	0.38	61.88	0.60
zh-en						
dev2010	9.53	0.38	33.96	0.37	81.71	0.95
tst2010	11.12	0.30	36.27	0.27	76.39	0.74
tst2011	13.34	0.37	38.77	0.32	65.91	0.41
en-fr						
dev2010	25.28	0.57	46.86	0.46	57.48	0.68
tst2010	28.46	0.49	49.14	0.38	51.69	0.47
tst2011	33.74	0.71	53.68	0.52	44.83	0.61

Table 6: Performance of baselines in terms of %BLEU, METEOR (mtr) and TER scores; σ stands for standard deviation. Values were computed in case sensitive mode and taking into account punctuation marks.

Although models were strictly trained on in-domain data and a quite standard configuration of Moses was used for both training and running translations, results on BLEU and TER compare well with those obtained on `tst2011` by participants at the MT track (Federico et al., 2011), whose ranges are summarized in Table 7. METEOR values seem not to be comparable, likely due to a different setup of the language dependent modules of the scorers.

tst2011	%bleu	mtr	ter
ar-en	19.56–26.32	54.66–61.10	64.65–55.81
zh-en	11.90–16.89	45.91–52.84	70.66–62.80
en-fr	34.39–37.65	24.46–27.14	45.69–41.70

Table 7: Ranges of official scores ("case sensitive+punctuation" mode) obtained by IWSLT 2011 evaluation campaign participants on the evaluation set `tst2011`.

4.2 New Language Pairs

Four new language pairs taken from Table 1 have been here considered, namely Dutch–to–English, German–to–English, German–to–Italian and English–to–Italian. These pairs as a whole cover many interesting issues: translation involving inflected languages at different extent (German, Italian, Dutch), compound words (German, Dutch), translation between non-English languages (German–to–Italian), among others. Moreover, in three cases out of four, the amount of available parallel training data is of the order of 1 million words.

4.2.1 Data

Texts used for these experiments are available at the WIT^3 website. The same talks defining the IWSLT `dev2010` and `tst2010` sets were used for tuning and evaluation purposes, respectively. The rest of parallel data was used for training translation and reordering models. LMs were estimated on all talks available for each target language excluding the talks of development and test sets. Tables 8 and 9 show some statistics of collected texts after tokenization.

4.2.2 Performance

Baselines were developed exactly as for the IWSLT 2011 language pairs. Table 10 provides performance on both the tuning set `dev2010` and the evaluation set `tst2010`. In order to assess the quality of our baseline systems only on

text	#sent.	Dutch		English	
		$\|W\|$	$\|V\|$	$\|W\|$	$\|V\|$
parallel	54.6k	978k	46.0k	1.04M	32.7k
dev2010	932	18.1k	3.8k	20.2k	3.4k
tst2010	1367	24.7k	3.9k	26.2k	3.4k
text	#sent.	German		English	
		$\|W\|$	$\|V\|$	$\|W\|$	$\|V\|$
parallel	63.9k	1.16M	63.1k	1.22M	35.5k
dev2010	930	19.1k	4.2k	20.2k	3.4k
tst2010	1660	30.3k	5.2k	32.0k	3.9k
text	#sent.	German		Italian	
		$\|W\|$	$\|V\|$	$\|W\|$	$\|V\|$
parallel	56.1k	1.06M	59.8k	1.03k	48.9k
dev2010	886	18.4k	4.1k	17.1k	4.0k
tst2010	1597	30.3k	5.2k	29.3k	5.2k
text	#sent.	English		Italian	
		$\|W\|$	$\|V\|$	$\|W\|$	$\|V\|$
parallel	98.1k	1.95M	45.5k	1.80M	65.9k
dev2010	887	19.5k	3.3k	17.1k	4.0k
tst2010	1598	32.0k	3.9k	29.3k	5.2k

Table 8: Statistics on parallel data used for setting up the baselines on additional language pairs. See Table 4 for the meaning of symbols.

monolingual	#sent.	$\|W\|$	$\|V\|$
English	128.3k	2.49M	51.5k
Italian	100.8k	1.85M	67.0k

Table 9: Statistics on monolingual data used for training LMs of additional baselines. See caption of Table 4 for the meaning of symbols.

the basis of their automatic scores, we leverage the large-scale investigation reported in (Coughlin, 2011) where translations judged as acceptable or at least almost acceptable in human evaluations corresponded to %BLEU scores ranging in 20–30. Hence, our baselines provide good translations towards English, despite the quite limited amount of available parallel training data, and adequate for the English–to–Italian pair. On the contrary, the German–to–Italian direction turns out to be more difficult: this could be due either to the scarcity of training data or to the inadequacy of German preprocessing (no word decompounding), or both.

5 WIT^3 Website

The WIT^3 website address is:

`http://wit3.fbk.eu`

	%bleu	σ	mtr	σ	ter	σ
nl-en						
dev2010	23.31	0.63	46.63	0.48	57.96	0.63
tst2010	30.99	0.53	54.47	0.36	48.75	0.51
de-en						
dev2010	26.71	0.58	51.89	0.39	50.86	0.56
tst2010	25.88	0.46	50.57	0.34	52.13	0.47
de-it						
dev2010	13.17	0.46	28.65	0.48	69.89	0.57
tst2010	13.06	0.34	28.59	0.37	68.87	0.42
en-it						
dev2010	22.43	0.58	39.16	0.55	57.61	0.60
tst2010	22.14	0.42	39.44	0.41	56.08	0.44

Table 10: Performance of baselines on additional language pairs in terms of %BLEU, METEOR (mtr) and TER scores; σ stands for standard deviation. Values were computed in case sensitive mode and taking into account punctuation marks.

The website currently hosts the TED Talks Corpus. We expect to include other collections of talks in the future, too. Concerning the TED Talks, the corpus version will be updated on a regular basis as soon as new translations will be from the original site. For each, version the following information will be available:

XML: the set of XML files with all talks subtitled in each language

Parallel: an active web page resembling Table 1; each entry links to an archive including parallel text for training and, if any, for development and evaluation purposes

DTD: the schema defining the XML format used for storing TED talks.

The website provides the following software tools, too:

`find-common-talks.pl`: given the XML files of TED talks in two languages, it outputs the set of `talkid`'s (see Sections 2.1 and 2.2) for which subtitles are available in both those languages

`filter-talks.pl`: it selects from a given XML file the talks whose `id`'s are passed as parameter

`ted-extract-par.pl`: given a pair of XML files, it extracts the text from the `transcript` field (Sections 2.1, 2.2) of common talks, aligned at the caption level

`ted-extract-mono.pl`: given an XML file, it extracts the text of talks from the `transcript` field (Sections 2.1, 2.2)

`rebuild-sent.pl`: it re-builds sentences from captions (Section 2.2).

By exploiting the XML files and the supplied tools, one can extract the set of common talks for each possible language pair, as well as the monolingual text.

For many language pairs, the site will already provide training, development, and evaluation data sets, while for others, only the parallel text.

It is worth noticing that the `url` tag (see Section 2.1) allows the retrieval of the original HTML document of each talk, this way giving the possibility to users to build from scratch their own linguistic resource based on TED talks.

6 Summary

In this paper, we have described WIT^3, a web inventory distributing the multilingual subtitles available under the TED Talks website. We believe, this collection represents a precious resource for the MT community given its size and its variety in terms of both languages and topics covered. In fact, more than 900 talks had been collected at the end of 2011, subtitled in up to 82 languages and spanning the whole of human knowledge.

We hope WIT^3 will offer an adequate service to the research community by distributing: (i) parallel texts, benchmarks and reference MT results for some language pairs; and (ii) original formatted files and tools for processing them to let anyone build his/her own data sets for any language pair.

Acknowledgments

This work was supported by the EU-Bridge project (FP7-ICT-2011-7), which is funded by the EC under the 7^{th} Framework Programme.

The authors also thank the three anonymous reviewers for their helpful comments.

References

Callison-Burch, C., P. Koehn, C. Monz, and O. Zaidan 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proc. of WMT*, pp. 22-64, Edinburgh, Scotland, UK.

Chen, S. F. and J. Goodman. 1999. An Empirical Study of Smoothing Techniques for Language Modeling. *Computer Speech and Language*, 4(13):359–393.

Clark, J., C. Dyer, A. Lavie, and N. Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proc. of ACL*, Portland, US-OR.

Coughlin, D. A. 2003. Correlating Automated and Human Assessments of Machine Translation Quality. In *Proc. of MT Summit*, pp. 23-27, New Orleans, US-LA.

Diab, M., K. Hacioglu, and D. Jurafsky. 2004. Automatic Tagging of Arabic Text: from Raw Text to Base Phrase Chunks. In *Proc. of HLT-NAACL: Short Papers*, pp. 149–152, Boston, US-MA.

Eisele, A. and Y. Chen. 2010. MULTIUN: A Multilingual Corpus from United Nation Documents. In *Proc. of LREC*, Valletta, Malta.

Federico, M., N. Bertoldi, and M. Cettolo. 2008. IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In *Proc. of Interspeech*, pp. 1618–1621, Melbourne, Australia.

Federico, M., L. Bentivogli, M. Paul, and S. Stüker. 2011. Overview of the IWSLT 2011 Evaluation Campaign. In *Proc. of IWSLT*, San Francisco, US-CA.

Girardi. 2011. The HLT Web Manager. *FBK Technical Report n. 23969*. Trento, Italy. `https://wit3.fbk.eu/tools/WebManager_Manual.pdf`

Koehn, P. et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of ACL: Demo and Poster Sessions*, pp. 177–180, Prague, Czech Republic.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. of MT Summit*, pp. 79–86, Phuket, Thailand.

Paul, M., M. Federico, and S. Stücker. 2010. Overview of the IWSLT 2010 Evaluation Campaign. In *Proc. of IWSLT*, pp. 3–27, Paris, France.

Takezawa, T., G. Kikui, M. Mizushima, and E. Sumita. 2007. Multilingual Spoken Language Corpus Development for Communication Research. *Computational Linguistics and Chinese Language Processing*, 12(3):303–324.

Tiedemann, J. 2009. News from OPUS - a Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Recent Advances in Natural Language Processing (vol V)*, pp. 237–248. John Benjamins, Amsterdam/Philadelphia.

Tseng, H., P. Chang, G. Andrew, D. Jurafsky, and C. Manning. 2005. A Conditional Random Field Word Segmenter. In *Proc. of SIGHAN Workshop on Chinese Language Processing*, Jeju Island, Korea.