Going beyond zero-shot MT: combining phonological, morphological and semantic factors. The UdS-DFKI System at IWSLT 2017

Cristina España-Bonet and Josef van Genabith

University of Saarland and DFKI, Saarbrücken, Germany

`{cristinae,Josef.Van_Genabith}@dfki.de`

Abstract

This paper describes the UdS-DFKI participation to the multilingual task of the IWSLT Evaluation 2017. Our approach is based on factored multilingual neural translation systems following the small data and zero-shot training conditions. Our systems are designed to fully exploit multilinguality by including factors that increase the number of common elements among languages such as phonetic coarse encodings and synsets, besides shallow part-of-speech tags, stems and lemmas. Document level information is also considered by including the topic of every document. This approach improves a baseline without any additional factor for all the language pairs and even allows beyond-zero-shot translation. That is, the translation from unseen languages is possible thanks to the common elements —especially synsets in our models— among languages.

1. Introduction

Neural machine translation systems (NMT) are currently the state of the art for most language pairs [1] and, among other advantages with respect to other paradigms, they can be easily extended to multilingual systems (ML-NMT) [2, 3]. ML-NMT systems usually use a common vocabulary where some words are shared and, more importantly, they project all the languages into the same embedding space clustering sentences according to their meanings. However, the clustering is not perfect and especially distant languages or those with fewer data are more difficult to group by semantics [4].

With the aim of facilitating the semantic clustering of languages, we enrich words with several levels of annotation. The highest level of annotation is represented by *Babel synsets*. BabelNet (BN) is a multilingual semantic network connecting concepts via synsets [5]. Each concept, or word, is identified by its ID irrespective of its language, effectively turning these IDs interlingua. At a lower level, we start from the premise that languages, especially within families, share roots that have evolved with time. We use stems and lemmas to capture common roots and phonetic coarse encodings for phonetic similarities.

On the other hand, we also take advantage of the coherent structure of the training data composed by a collection of TED talk transcriptions in several languages. We inform the system about the topic of every word according to the document it belongs to, expecting to improve lexical selection in this way. Previous research modifies a standard encoder-decoder architecture to deal with extra-sentence information [6, 7]. Here we take the opposite approach and modify (annotate) the data in order to capture relevant knowledge.

Technically, we include all the aforementioned information as factors in a ML-NMT system. Following [8], each feature has its own word vector which is concatenated to the BPE'd token vectors to build the hidden states. So, when including factors in the source representation of a token, every source token embedding has the top-k elements describing the distribution of the word and the remaining elements describing other features.

This is not the first time that linguistic factors are used in NMT. We use the implementation in [8], but also the authors in [9, 10] use part of speech tags and grammatical information in their systems. However, to our knowledge, this is the first time that factors are designed to try to reduce distances between different source languages and therefore to mimic the effect of having a larger corpus. Also, BPE subunits are expected to be more descriptive since semantic information of the complete word is added to its representation. A vocabulary expansion is naturally produced by the concatenation of the different features. Our approach specifically targets multi- and interliguality in translation and, as an extreme effect, we show how *beyond-zero-shot translation* can be possible, that is, the translation from unseen source languages thanks to interlingual factors.

The rest of the paper is organised as follows. Section 2 describes all the factors used in this work and which tools and methodologies we use to obtain them. In Section 3 we analyse the characteristics of the training corpus with respect to these factors. Section 4 briefly describes the parameters of the ML-NMT systems and Section 5 reports the results in the small data and zero-shot training conditions. Finally, we summarise and draw our conclusions in Section 6.

2. Linguistic and Semantic Annotations

Coarse-Grained Part of Speech (p). We use a coarse-grained part-of-speech (PoS) tag set with 10 elements: {`NOUN`, `VERB`, `PREPOSITION`, `PRONOUN`, `DETERMINER`, `ADVERB`, `ADJECTIVE`,

Table 1: Statistics for the number of elements in the monolingual TED corpora (*all*, top block; *unique*, bottom block). Monolingual corpora have been built as the concatenation of all the parallel counterparts eliminating duplicates.

	West Germanic Languages			Latin Languages			
	en	*de*	*nl*	*ro*	*it*	*es*	*fr*
Sentences	545,270	303,668	444,287	225,980	513,693	151,631	140,717
Tokens	9,768,374	5,148,199	6,894,438	3,732,679	8,367,940	2,494,336	2,473,040
uToken	141,013	221,459	187,148	213,670	200,697	148,366	131,015
uLemma	73,048	101,003	85,846	72,535	52,525	52,052	53,088
uStem	50,128	94,126	85,560	54,227	44,691	35,307	40,504
uM3	57,630	79,029	60,534	30,576	32,828	31,840	32,234
uBN	28,445	34,022	27,720	24,375	27,172	23,567	23,856

Table 2: Coverage of the different subcorpora (in %) by the common elements among the languages. The number in parenthesis shows the absolute number of common elements.

	Germanic	Latin	ALL	SMALL	ZERO
Token	40% (17,185)	32% (11,690)	30% (8,279)	30% (13,150)	32% (13,150)
Lemma	56% (14,576)	37% (9,096)	40% (7,922)	41% (11,835)	43% (11,835)
Stem	57% (12,029)	52% (8,114)	46% (4,971)	45% (7,452)	47% (7,452)
M3	87% (9,961)	87% (8,164)	84% (5,922)	69% (7,506)	70% (7,506)
BN	15% (5,507)	27% (6,104)	12% (2,367)	12% (3,291)	12% (3,291)

CONJUNCTION, ARTICLE, INTERJECTION}. This tag set is defined so as to be compatible with the one in the BabelNet ontology, so this set does not exactly correspond to the Universal Part-of-Speech Tagset [11] although the granularity is similar. We use the IXA pipeline [12] to annotate English, German, Spanish and French documents with PoS and TreeTagger [13] for Dutch, Romanian and Italian. The original tags are then mapped to our common reduced tagset[1].

Lemma (l). As with PoS, we use the IXA pipeline for English, German, Spanish and French; and TreeTagger for Dutch, Romanian and Italian.

Stem (s). Stems are obtained with the Snowball API which implements the Porter algorithm [14].

Approximate Phonetic Encoding (m). We use a phonetic algorithm to encode words by their pronunciation. The purpose is to bring close languages together by taking advantage of similar pronunciations in a similar way lemmas and stems do for close spellings. Phonetic algorithms that provide a coarse encoding of a word are more appropriate for this task than the real phonetic transcription which would be too discriminative.

Phonetic algorithms like Soundex [15] or Metaphone [16] are usually developed for a particular language with possible adjustments to deal with specific features of another one such as matching names. As an approximation, in our experiments we use Metaphone 3 for English on all the languages. Metaphone 3 (M3) is a phonetic algorithm that takes into account irregularities in English coming from several languages including Germanic and Latin ones. As generic features, the encoding converts all the initial vowels into an A and pairs of unvoiced and voiced consonants are encoded by the same letter. The algorithm is commercially available also for Spanish and German, but the only open source resource that we know of is for the English version[2].

Babel Synset (b). BabelNet [5] is a multilingual semantic network connecting concepts via *Babel synsets*. We enrich TED data content words with their synset information. For this, we select (*i*) nouns (including named entities, foreign words and numerals), (*ii*) adjectives, (*iii*) adverbs and (*iv*) verbs following the mappings to our coarse-grained part-of-speech tags. In addition, we explicitly mark negation particles with a tag NEG and include them here to account for their semantics.

A word can have several Babel synsets. We retrieve a synset according to the lemma and PoS of a word. In case there is still ambiguity, as it is in most of the cases, we select the BabelNet ID as the first ID according to its own sorting: *(a) puts WordNet synsets first; (b) sorts WordNet synsets based on the sense number of a specific input word; (c) sorts Wikipedia synsets lexicographically based on their main sense.*

Topic (t). TED talks are tagged with a set of English keywords that describe the topic of a document. Topic information can be relevant under two points of view: (*i*) given

[1] The mappings and the full annotation pipeline can be obtained here: https://github.com/cristinae/BabelWE

[2] Metaphone 3 is available within the OpenRefine tool, https://github.com/OpenRefine/OpenRefine

Table 3: Characterisation of the 20 topics learned with a BTM system. The percentage and absolute value of documents in the training corpus of every topic is shown together with the top-5 keywords that describe them.

Label	Proportion	Top-5 keywords
t1	10.6% (206)	science (6.5%) biology (5.6%) health (3.9%) medical research (3.9%) medicine (3.8%)
t2	10.0% (193)	culture (6.6%) entertainment (5.9%) technology (5.8%) design (5.3%) business (4.1%)
t3	8.2% (160)	culture (5.2%) entertainment (4.8%) art (4.0%) storytelling (3.1%) humor (3.1%)
t4	7.3% (141)	brain (7.3%) science (5.8%) neuroscience (5.3%) psychology (5.1%) mind (5.0%)
t5	6.6% (128)	global issues (5.0%) future (3.9%) society (3.9%) government (3.7%) politics (3.6%)
t6	6.6% (127)	environment (6.0%) science (5.0%) ecology (4.2%) plants (4.1%) nature (3.9%)
t7	6.0% (116)	technology (9.2%) computers (5.5%) design (4.7%) Internet (3.5%) TEDx (3.1%)
t8	5.9% (113)	technology (6.0%) environment (5.5%) science (4.7%) sustainability (4.6%) global issues (4.6%)
t9	5.3% (102)	science (7.1%) animals (5.4%) environment (5.0%) oceans (4.6%) biodiversity (4.5%)
t10	4.7% (90)	global issues (10.9%) politics (6.8%) war (5.8%) culture (4.8%) TEDx (4.0%)
t11	4.2% (81)	design (9.5%) technology (8.6%) invention (7.5%) innovation (5.8%) creativity (4.4%)
t12	4.0% (78)	global issues (9.2%) business (9.0%) economics (6.9%) culture (5.6%) Africa (4.5%)
t13	4.0% (77)	science (9.8%) technology (6.5%) space (6.1%) universe (5.7%) astronomy (5.4%)
t14	3.9% (77)	health (10.6%) healthcare (8.9%) medicine (8.2%) science (6.5%) technology (4.8%)
t15	3.7% (72)	technology (7.0%) science (6.4%) biology (4.2%) design (3.7%) robots (3.7%)
t16	2.7% (53)	women (7.3%) social change (5.7%) culture (5.1%) education (5.0%) activism (5.0%)
t17	2.1% (40)	design (13.5%) cities (10.1%) architecture (8.1%) art (4.9%) infrastructure (4.2%)
t18	1.8% (35)	music (14.7%) performance (13.8%) entertainment (12.9%) live music (10.2%) piano (3.6%)
t19	1.2% (24)	work (7.5%) business (5.6%) motivation (5.4%) personal growth (5.3%) success (4.4%)
t20	1.2% (23)	culture (12.9%) religion (9.5%) global issues (8.0%) philosophy (5.6%) science (5.1%)

a document, it is shared across languages, so it can help the NMT system to locate together in the embedding space the same sentence across languages, and (*ii*) it may improve document-level translation since it can help to disambiguate word translations according to its topic.

With a total of 390 different keywords and a mean of 6.5 per document, considering all of them as input information for the NMT system would lead to too much diversity. Besides, some keywords such as *technology*, *science*, *culture* or *global issues* are very frequent and could put in irrelevant information. Therefore, we decided to learn a topic model on the keywords and tag each document with a single interlingua label. Since a document is then only the short set of keywords in English, we apply a monolingual biterm topic model (BTM) for short texts [17] for the purpose.

As an alternative, we also apply polylingual topic models learned with Mallet [18] on all documents using the full vocabulary. However, after inferring the topic of each document, we obtained a mixture of top-k topics that did not allow a unique labelling of the same document across languages and the use of a single label would not be an interlingua tag as desired. Since keywords are always available for TED talks we used the first approach.

3. Corpus Characteristics

We use the corpus provided for the IWSLT 2017 multilingual task [19]. It comprises transcripts and manual translations of the TED talks accessible on April 26th, 2017. Two sets, *dev2010* and *tst2010*, are available for validation and testing purposes. The corpus includes documents in five languages, *en-de-ro-it-nl*, summing up to 9161 talks. The intersection of talks among languages is high, 7945 documents are common to all of them. In addition, we also use TED talks in French and Spanish obtained from previous IWSLT campaigns[3]. This data is not used for training, but we include them in the analysis of the corpus because in a subsequent section we explore the translation from unseen languages.

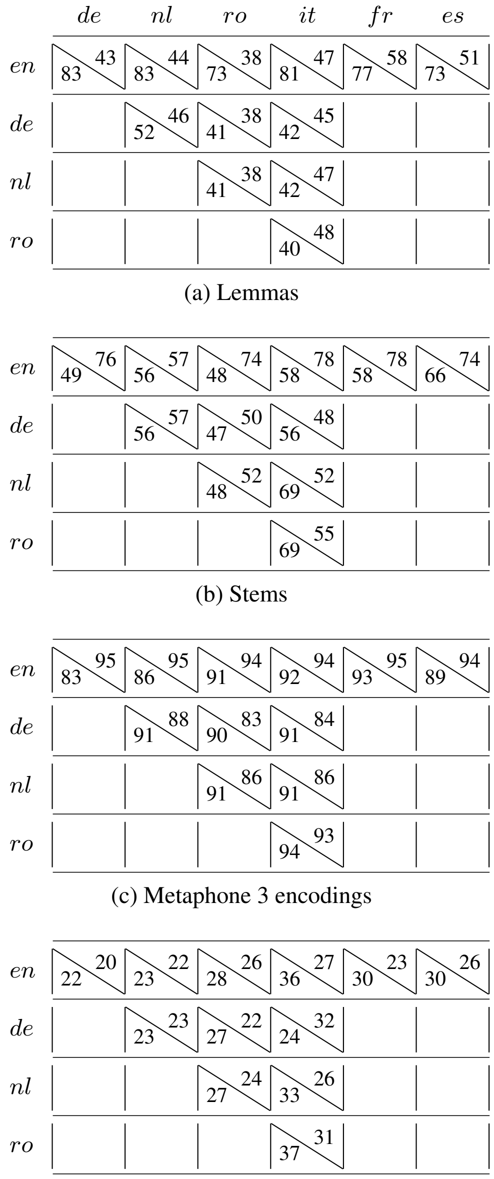
Table 1 shows the general statistics of the TED corpus by language. Languages are divided into two families: West Germanic with en, de and nl, and Latin with ro, it, es and fr. Notice that en, ro and it have significantly more sentences and that could benefit the translation from/to these languages, but the number of unique tokens ($uToken$) is quite homogeneous with the exception of fr and es.

The number of unique elements in the corpus decreases when going from words, to lemmas, stems, M3 encodings and BN synsets. The only exception is en, where we obtain more unique M3 encodings than stems. The number of unique elements is an indication of the ambiguity given by the factor: words are the least ambiguous linguistic factor but too many to be fully covered by the vocabulary of ML-NMT systems, and M3 encodings are the most ambiguous elements up to the point that they frequently erase the differences between unrelated words. In English, `anyone` and `union` share the same M3 encoding `ANN` but not the meaning. The same encoding applies to the German words `eine` and `ihnen` or the Italian ones `unione` or `annoiano`, some of them are translations, some of them not. BN synsets are not directly comparable because they are only obtained for a subset of PoS tags.

Our main interest is to observe the intersection of these elements in different languages. Table 2 reports the percentage of a corpus that is covered by the common elements among all the languages that build up such corpus. We show these figures for five corpora: Germanic including en, de and

[3] `https://wit3.fbk.eu/mt.php?release=2014-01`

Figure 1: Percentage of TED corpora covered by the common elements in a language pair. A cell represents the language pair row–column, with the coverage of row language given by the bottom subcell and the coverage of the column language given by the upper subcell.



	de	nl	ro	it	fr	es
en	43 83	44 83	38 73	47 81	58 77	51 73
de		46 52	38 41	45 42		
nl			38 41	47 42		
ro				48 40		

(a) Lemmas

	de	nl	ro	it	fr	es
en	76 49	57 56	74 48	78 58	78 58	74 66
de		57 56	50 47	48 56		
nl			52 48	52 69		
ro				55 69		

(b) Stems

	de	nl	ro	it	fr	es
en	95 83	95 86	94 91	94 92	95 93	94 89
de		88 91	83 90	84 91		
nl			86 91	86 91		
ro				93 94		

(c) Metaphone 3 encodings

	de	nl	ro	it	fr	es
en	20 22	22 23	26 28	27 36	23 30	26 30
de		23 23	22 27	32 24		
nl			24 27	26 33		
ro				31 37		

(d) Babel synsets

nl; Latin with ro, it, es and fr; ALL with the sum of Germanic and Latin; and SMALL and ZERO with the languages considered for the multilingual translation task en, de, nl, ro and it. In general, Germanic languages share more vocabulary (tokens, lemmas and stems) than Latin languages; the disparity in lemmas is more marked in Latin languages: whereas $9,096$ common lemmas cover only a 37% of the corpus, $8,114$ common stems cover a 52% of it. It is remarkable to notice the percentage of common vocabulary in the ALL corpus (30% for tokens, 40% for lemmas and 46% for stems). These high values justify their usage in multilingual systems.

M3 encodings clearly show an excess of ambiguity: 87% of the Germanic and Latin corpora are covered by the common encodings, 70% of the SMALL and ZERO ones. Still, since the information is complementary to the previous elements, we employ it in the translation systems. Finally, the percentage of common BN synsets is higher for the Romance languages (27% vs. 15%). Joining all the languages together decreases this to 12%. Differently to the other factors, BN synsets only cover 4 out of the 10 PoS tags. Besides, they suffer from a *sense effect*: whereas `investigación` in Spanish and `investigation` in English share stem and M3 encoding, the top BabelNet ID is `bn:00067280n` for Spanish and `bn:00047355n` for English because the first sense of the word in the two languages is different.

Figure 1 shows the equivalent analysis per language pair. Notice that the English corpus is the best covered by common lemmas, stems and M3 encodings and that differences between languages can be large, especially when English is involved. According to these numbers, this is the language least rich in lemmas, stems and diversity of pronunciations.

Finally, we analyse the data according to their theme. To do so, we infer the most probable topic for each document with a BTM model learned for 20 topics, so that each topic is the main topic of at least 1% of the training documents. Table 3 shows the characterisation of each topic and the percentage of the corpus described by them. Note that although the extracted topics define different themes, they share keywords. In other words, the diversity in the TED talks is low and themes are close to each other.

4. NMT Systems

Our system is a many-to-many NMT engine trained with Nematus [20]. As done in [3] and similarly to [2], the engine is trained on parallel corpora for the several language pairs simultaneously, 16 pairs for the zero-shot training condition (ZERO) and 20 for the small data training condition (SMALL), with the only addition of a tag in the source sentence to account for the target language "<2trg>". SMALL includes all the pairs generated from the en-de-ro-it-nl languages and ZERO excludes the de-nl and it-ro pairs. In both cases, we only consider those sentences with less than 50 tokens for training, that is 2.113.917 parallel sentences (39.393.037 tokens) in the first case, 1.692.594 sentences (31.671.455 tokens) in the second one.

We consider each token in a source sentence to be represented by (a subset of) the features introduced in the previous sections. The final representation of a word is the concatenation of all its features. This has been named *factored translation* by their similarities with factored translation in statistical machine translation [21] and we use the implementation available in Nematus [8]. The same work [8] explores the inclusion of PoS and subword tags, morphological features, lemmas and syntactic dependency labels as input features for bilingual NMT systems involving, en, de and ro. Here, we

Table 4: Dimensions per factor in the final word embedding for the systems shown in the most-left column.

	token w	PoS p	lemma l	stem s	M3 m	BN b	topic t
w	506	0	0	0	0	0	0
wl	300	0	206	0	0	0	0
ws	300	0	0	206	0	0	0
wm	300	0	0	0	206	0	0
wb	300	0	0	0	0	206	0
wt	496	0	0	0	0	0	10
$wpsm$	300	6	0	100	100	0	0
$wpsmb$	275	6	0	75	75	75	0
$wpsmt$	290	6	0	100	100	0	10
$wpsmbt$	265	6	0	75	75	75	10

extend the model to use more generic factors such as stems and M3 encodings, and interlingual factors such as Babel synsets. The next example shows a truecased phrase annotated with token|PoS|stem|M3|BN in English and German:

en: that|DETERMINER|that|0T|-
's|***VERB***|'s|S|bn:00083181v the|***DETERMINER***|the|0|-
problem|***NOUN***|problem|***PRPLM***|***bn:00048242n***
de: das|PRONOUN|das|TS|-
ist|***VERB***|ist|AST|- das|***DETERMINER***|das|TS|-
Problem|***NOUN***|probl|***PRPLM***|***bn:00048242n***

where boldface emphasises differences and boldface plus italics emphasises similarities. The examples belong to two close languages that share vocabulary and roots. However, problem and Problem would not match without the information on PoS, M3 encoding and BN synset. The example displays other characteristics such as differences of PoS between languages (DETERMINER vs. PRONOUN) for that/das, or lacking BN synset in a language. Differences in the retrieved BN sense are not seen here but should also be considered (portrait|NOUN|portrait|PRTRT|bn:00063682n vs. Porträt|NOUN|porträt|PRTRT|bn:00063683n).

All our systems employ a common vocabulary of 150 K tokens plus 2 K for subword units segmented using Byte Pair Encoding (BPE) [22]. Subwords in the source sentence are annotated with the same factors as the complete word has. As for the parameters, we use a learning rate of 0.0001, Adadelta optimisation, 800 hidden units, a mini-batch size of 100, and drop-out only for hidden layers and input embeddings. We also tie the embeddings in the decoder side to reduce the size of the translation models. The dimension of the word embeddings is always 506, but every model has a different distribution of the dimensions per factor. We refer the reader to Table 4 to see the distribution, where models are named using the letters that represent the factors included.

5. Results and Discussion

Below we report the translation performance for several systems under the small data and zero-shot training conditions. We evaluate systems that combine word tokens (w) with the individual linguistic or semantic factors (wp, wl, ws, wm, wb and wt) and the combination of additional factors ($wpsm$, $wpsmb$, $wpsmt$ and $wpsmbt$). As BabelNet was not within the allowed resources, our submissions for both training conditions were: $wpsm$ (primary, SUB1) and $wpsmt$ (SUB2) and $wpsmbt$ (SUB3) as contrastive.

Results are broken down according to the training condition and language pair: Table 5 shows the BLEU scores on truecased and tokenised translations under the zero-shot training condition and Table 6 shows the equivalent under the small one. First of all, we obtain the results for three different decoding settings on our baseline with only words: two beam sizes, 5 ($w5$) and 10 ($w10$); and an ensemble with the last four models with a beam size of 10 (w). Increasing the beam size is the major source of improvement (1.5 BLEU points on the concatenated test set), and this number is further increased by the ensemble up to 2.4 BLEU points. We analyse the effect of the designed factors over this strong baseline. Since conclusions are analogous, the most detailed analysis is only reported for the zero-shot training condition.

Notice that the global BLEU score for SMALL systems is better than for ZERO mainly because of the zero-shot pairs de-nl and it-ro. For the other pairs, the enlargement of the multilingual corpus is even harmful both in a baseline with only words and with factored models. When considering the performance of the systems on all the languages simultaneously, the best system is the one exploiting all the features ($wpsmbt$), with a BLEU of 25.46 for ZERO and 25.72 for SMALL. These scores are close to but below our primary submission (25.38 for ZERO and 25.70 for SMALL) which does not consider BN synsets or topic labels.

In general and for most language pairs, BN synsets are the only factor that is able to produce translation improvements by itself, the other ones are in average below the baseline but help to break degeneracies when combined and produce a beneficial effect. PoS tags also achieve a small improvement, but it is non-significant and much less than the one obtained by the authors in [8, 9] for bilingual NMT systems. Stems and lemmas perform equally well in average with only few exceptions: stems are better for translating from de or into nl, while lemmas are better for translating into de. For other language pairs differences are either non-systematic or insignificant. M3 encodings alone are too ambiguous as shown by the high percentage of the corpus covered by common encodings already at the bilingual level (see Figure 1c). Note that in the case of de, where the percentage is lower, the encodings do help to increase the performance. As expected, topic information does not imply relevant changes probably due to the low diversity in the topic characteristics (Table 3). However, the fact that contrary to previous research [8, 9, 10] neither PoS tags nor lemmas have a positive impact in the ML-NMT system motivates further experiments with bilingual NMT systems enriched with M3 encodings, BN synsets and topic information.

Figure 2: 2D t-SNE representation of the context vectors of the first 8 source sentences of *tst2010* for system w, wb and $wpsmb$ under the zero-shot training condition. The same sentence has the same colour in different languages.

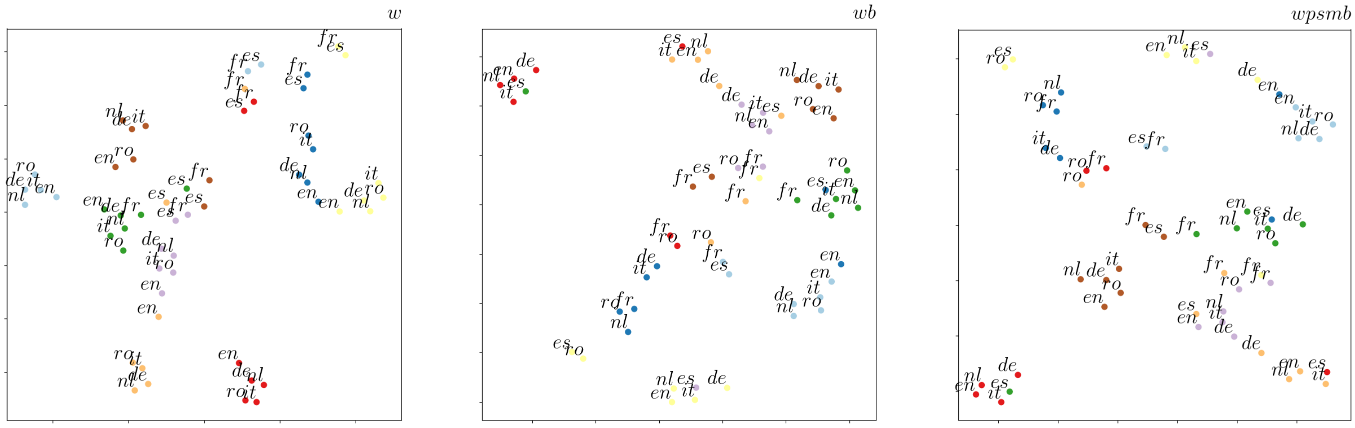


Table 5: BLEU scores on the TED talks *tst2010* obtained with several systems under the zero-shot training condition. The zero-shot pairs, *de-nl* and *it-ro*, are shown at the end. `SUB1`, `SUB2` and `SUB3` were submitted to the shared task.

	beam size		factors + 4-ensembles (beam size 10)										
	$w5$	$w10$	w	wp	wl	ws	wm	wb	wt	$wpsmb$	$wpsm$ (`SUB1`)	$wpsmt$ (`SUB2`)	$wpsmbt$ (`SUB3`)
de2it	18.02	19.20	19.78	19.85	18.95	20.07	19.67	20.28	19.67	**20.35**	20.10	20.05	20.33
it2de	18.05	19.49	19.90	20.03	20.03	19.98	19.98	20.42	20.30	20.22	20.42	20.06	**20.45**
de2ro	15.85	17.57	18.23	17.98	17.16	17.73	17.96	18.46	18.19	**18.60**	18.23	18.00	18.40
ro2de	18.56	20.05	20.87	21.04	**21.52**	21.06	20.93	21.23	20.78	21.34	21.49	21.12	21.41
de2en	30.11	31.67	32.65	32.62	31.66	32.74	32.47	32.97	32.71	33.34	33.11	32.91	**33.51**
en2de	24.61	26.06	27.02	**27.53**	27.30	26.80	27.10	27.26	26.97	27.36	27.15	27.10	27.44
en2it	26.33	27.90	28.88	28.66	28.74	28.97	28.41	**29.35**	28.69	29.06	28.99	28.94	29.34
it2en	31.22	32.56	33.46	33.59	33.85	33.15	32.95	33.20	33.25	33.49	33.53	33.33	**33.87**
en2nl	28.60	30.24	31.27	31.21	31.12	30.87	30.85	31.08	31.26	30.80	30.90	31.17	**31.44**
nl2en	33.86	35.39	36.20	36.61	36.34	36.56	36.16	36.57	36.03	36.92	36.82	36.55	**37.40**
en2ro	23.65	25.28	26.38	26.37	25.67	25.83	25.19	26.18	25.76	26.37	25.85	26.08	**26.47**
ro2en	32.02	33.59	34.34	34.33	34.60	34.40	34.28	34.82	34.34	**35.31**	34.87	34.89	35.09
it2nl	19.03	21.05	21.58	21.65	21.65	21.23	21.25	**21.91**	21.48	21.41	21.79	21.77	21.54
nl2it	19.80	21.23	21.72	21.56	21.34	21.62	21.16	**21.97**	21.71	21.81	21.61	21.84	21.83
nl2ro	17.28	18.42	19.09	18.89	18.98	18.69	18.78	19.39	19.07	19.35	19.09	**19.45**	19.42
ro2nl	19.28	21.21	21.70	21.72	21.76	21.79	21.74	21.65	22.00	22.21	**22.61**	22.20	22.50
de2nl	18.82	21.11	21.75	21.58	20.78	21.76	21.66	22.51	21.62	21.73	**22.29**	22.10	21.90
nl2de	18.82	20.76	21.52	21.81	21.86	21.46	21.62	21.99	21.56	**22.04**	21.81	21.99	21.77
it2ro	16.42	18.14	19.16	19.06	18.94	18.47	18.59	18.94	18.68	**19.51**	19.29	19.13	18.73
ro2it	17.37	19.50	20.04	20.17	20.61	20.38	19.97	20.84	20.28	20.60	**20.94**	20.74	20.32
Concatenation	22.68	24.31	25.08	25.10	24.93	24.96	24.82	25.32	25.01	25.38	25.33	25.30	**25.46**

It is interesting to notice that the final effect of the most *interlingual* factors has not been a better clustering of sentences according to their meaning. Figure 2 shows a qualitative example using a 2D t-SNE representation [23] of the context vectors of 8 sentences in 3 cases. The baseline ML-NMT system w (most-left plot) does already a very good job in locating the sentences in consonance with their semantics. The sentences for the languages used in training lie together for the different languages, while sentences in the unknown languages fr and es group in two specific regions of the space irrespective of their meaning. The effect of BN synsets (middle plot) and M3 encodings (not shown in Figure 2) is to locate fr and es sentences close to the the other Latin languages ro and/or it. By looking at the examples, that means that similarities of the M3 encodings across close languages are too strong to be compared with the most distant languages, and that the top-1 BN synset for a term usually depends on the family that the language belongs to. So,

Table 6: BLEU scores on the TED talks *tst2010* obtained with several systems under the small data training condition. `SUB1`, `SUB2` and `SUB3` were submitted to the shared task.

	wZERO	wSMALL	wpsm (SUB1)	wpsmt (SUB2)	wpsmbt (SUB3)
de2it	19.78	19.55	20.53	20.14	**20.58**
it2de	19.90	19.92	20.05	19.49	**20.26**
de2ro	18.23	18.07	18.21	**18.45**	18.05
ro2de	20.87	20.82	21.13	20.51	**21.33**
de2en	32.65	32.08	**33.44**	32.71	33.24
en2de	27.02	26.82	27.22	26.71	**27.37**
en2it	28.88	28.83	29.01	28.76	**29.07**
it2en	33.46	33.03	33.81	33.70	**33.85**
en2nl	31.27	30.72	31.10	31.02	**31.39**
nl2en	36.20	35.90	**37.00**	36.48	36.79
en2ro	**26.38**	25.57	26.09	25.86	25.99
ro2en	34.34	33.86	34.82	34.58	**34.89**
it2nl	**21.58**	21.16	21.36	21.30	21.49
nl2it	21.72	21.27	**21.82**	21.56	21.72
nl2ro	19.09	18.87	19.14	**19.35**	19.16
ro2nl	21.70	21.74	21.89	21.61	**22.27**
de2nl	21.75	22.97	23.67	**23.90**	23.46
nl2de	21.52	23.19	**23.92**	23.64	23.56
it2ro	19.16	20.31	**20.84**	20.79	20.67
ro2it	20.04	22.41	23.36	22.94	**23.70**
Concat.	25.08	25.12	25.70	25.50	**25.72**

the features designed in this way would maximise their effectiveness within a multilingual system for related languages and, at the light of current results, a better disambiguation and mapping between languages of synsets is necessary for a real interlingual setting. However, the current implementation already achieves statistically significant improvements when used in the *en-de-ro-it-nl*-NMT system and we show in the following how these features are useful to translate from unseen languages, *es* and *fr*. Translation into a new language is still not possible because the system cannot create new words beyond a combination of BPE subunits.

Table 7 summarises the results for *es*/*fr*–*en* translations using the multilingual system under the zero-shot training condition. When translating from English, the BLEU score is close to 1 for all system irrespective of the information they consider —also irrespective of the beam size an number of ensembled models. This score accounts mainly for the common words between the two languages. But, when translating into English, one can obtain a BLEU of 7.25 for *es*2*en* translation (5.07 for *fr*2*en*). The baseline is higher in this case because, as seen in Section 3, English is more sparse than the other languages. Even then, the baseline is improved by more than 4 points of BLEU for *es*2*en* and almost 3 points of BLEU for *fr*2*en*. The major contribution comes from the inclusion of Babel synsets (models *wb* and *wpsmb* outperform *wpsm*).

Table 7: BLEU scores for translations involving languages not seen at all in training, *es* and *fr*, on the *tst2010* under the zero-shot training condition.

	w	wp	wl	ws	wm	wpsm	wb	wpsmb
en2fr	1.11	**1.13**	1.05	0.98	0.98	1.00	1.04	1.04
fr2en	2.41	2.77	1.77	3.14	2.84	3.63	**5.07**	5.02
en2es	1.29	1.04	1.02	0.98	0.92	0.99	1.02	**1.36**
es2en	3.09	3.67	2.61	4.22	3.88	4.87	6.75	**7.25**

6. Conclusions

This paper describes the UdS-DFKI participation at IWSLT 2017. Besides the description of the engines, we analyse the multilingual TED corpus regarding our six characterisation factors: parts of speech, lemmas, stems, Metaphone 3 encodings, Babel synsets and topics.

The most promising feature turned to be BN synsets, especially when combined with other factors. However, our primary submission does not include them as the resource is not allowed in the small data training conditions. Our primary submission, the *wpsm* system, almost reaches the performance of our best system *wpsmbt* without any information on the topic and the sense of a token.

BN synsets are the most expensive factor to obtain and they are only queried for a subset of PoS; the common IDs cover between 20% and 40% of the parallel corpora, depending on the language pair. Even then, they improve translations for a 75% of the language pairs and allow beyond-zero-shot translation. Further efforts to deal with multiword expressions and resolve ambiguities in the retrieval of the synsets will be made to enhance the description of the data and facilitate a multilingual learning. Constraining other factors such as M3 encodings and topics to content words could also improve the performance and will be further researched.

7. Acknowledgements

This work was partially funded by the Leibniz Gemeinschaft (SAW-2016-ZPID-2), the BMBF through the project ALL SIDES (01IW14002) and the European Unions Horizon 2020 grant agreement No. 645452 (QT21).

8. References

[1] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, P. Koehn, V. Logacheva, C. Monz, M. Negri, M. Post, R. Rubino, L. Specia, and M. Turchi, "Findings of the 2017 Conference on Machine Translation," in *Proceedings of the Second Conference on Machine Translations (WMT 2017)*, September 2017, pp. 169–214.

[2] T. Ha, J. Niehues, and A. H. Waibel, "Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder," in *Proceedings of the International*

Workshop on Spoken Language Translation, Seattle, WA, November 2016.

[3] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. T. and1 Fernanda B. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation," *Transactions of the Association for Computational Linguist*, vol. 5, pp. 339–351, October 2017.

[4] C. España-Bonet, A. C. Varga, A. Barrón-Cedeño, and J. van Genabith, "An Empirical Analysis of NMT-Derived Interlingual Embeddings and their Use in Parallel Sentence Identification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1340–1350, December 2017.

[5] R. Navigli and S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artificial Intelligence*, vol. 193, pp. 217–250, 2012.

[6] L. Wang, Z. Tu, A. Way, and Q. Liu, "Exploiting Cross-Sentence Context for Neural Machine Translation," *CoRR*, vol. abs/1704.04347, 2017.

[7] J. Zhang, L. Li, A. Way, and Q. Liu, "Topic-Informed Neural Machine Translation," in *COLING*, 2016, pp. 1807–1817.

[8] R. Sennrich and B. Haddow, "Linguistic input features improve neural machine translation," in *Proceedings of the First Conference on Machine Translation*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 83–91.

[9] J. Niehues and E. Cho, "Exploiting linguistic resources for neural machine translation using multi-task learning," in *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark*, 2017, pp. 80–89.

[10] M. García-Martínez, L. Barrault, and F. Bougares, "Neural machine translation by generating multiple linguistic factors," in *Proceedings of the 5th International Conference on Statistical Language and Speech Processing (SLSP), Le Mans, France*, 2017, pp. 21–31.

[11] S. Petrov, D. Das, and R. McDonald, "A universal part-of-speech tagset," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.

[12] R. Agerri, J. Bermudez, and G. Rigau, "IXA pipeline: Efficient and Ready to Use Multilingual NLP Tools," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), may 2014.

[13] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *International Conference on New Methods in Language Processing*, Manchester, UK, 1994, pp. 44–49.

[14] M. F. Porter, "Snowball: A language for stemming algorithms."

[15] M. K. Odell, "The profit in records management," *Systems*, vol. 20, no. 20, 1956.

[16] L. Philips, "Hanging on the metaphone," *Computer Language Magazine*, vol. 7, no. 12, pp. 39–44, December 1990.

[17] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22Nd International Conference on World Wide Web*. New York, NY, USA: ACM, 2013, pp. 1445–1456.

[18] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit," 2002, http://mallet.cs.umass.edu.

[19] M. Cettolo, M. Federico, L. Bentivogli, J. Niehues, S. Stüker, K. Sudoh, K. Yoshino, and C. Federmann, "Overview of the IWSLT 2017 Evaluation Campaign," in *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT)*, Tokyo, Japan, 2017.

[20] R. Sennrich, O. Firat, K. Cho, A. Birch, B. Haddow, J. Hitschler, M. Junczys-Dowmunt, S. Läubli, A. V. Miceli Barone, J. Mokry, and M. Nadejde, "Nematus: a toolkit for neural machine translation," in *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain: Association for Computational Linguistics, April 2017, pp. 65–68.

[21] P. Koehn and H. Hoang, "Factored translation models," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 868–876.

[22] R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL, August 7-12, 2016, Berlin, Germany, Volume 1*, 2016.

[23] L. Van Der Maaten, "Accelerating t-SNE Using Tree-based Algorithms," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221–3245, January 2014.