Multilingual Semantic Networks for Data-driven Interlingua Seq2Seq Systems

Cristina España-Bonet and Josef van Genabith
Universität des Saarlandes and Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI)
Saarbrücken, Germany
{cristinae, Josef.Van_Genabith}@dfki.de

Abstract

Neural machine translation systems are state-of-the-art for most language pairs despite the fact that they are relatively recent and that because of this there is likely room for even further improvements. Here, we explore whether, and if so, to what extent, semantic networks can help improve NMT. In particular, we (*i*) study the contribution of the nodes of the semantic network, *synsets*, as factors in multilingual neural translation engines. We show that they improve a state-of-the-art baseline and that they facilitate the translation from languages that have not been seen at all in training (beyond zero-shot translation). Taking this idea to an extreme, we (*ii*) use synsets as the basic unit to encode the input and turn the source language into a data-driven interlingual language. This transformation boosts the performance of the neural system for unseen languages achieving an improvement of 4.9/6.3 and 8.2/8.7 points of BLEU/METEOR for $fr2en$ and $es2en$ respectively when neither corpora in fr or es has been used. In (*i*), the enhancement comes about because cross-language synsets help to cluster words by semantics irrespective of their language and to map the unknown words of a new language into the multilingual clusters. In (*ii*), because with the data-driven interlingua there is no *unknown* language if it is covered by the semantic network. However, non-content words are not represented in the semantic network, and a higher level of abstraction is still needed in order to go a step further and train these systems with only monolingual corpora for example.

Keywords: Multilingual Neural Machine Translation, Semantic Networks, BabelNet, Interlinguality

1. Introduction

The concept of semantic network was introduced by R.H. Richens in 1956 in relation to interlingual machine translation (IMT) (Richens, 1956). He defined a *semantic net* of *naked ideas* as what is left after removing the structural particularities of the base language. The elements of such a net represented things, qualities or relations. From 50 semantic primitives, Richens created the first semantic network, Nude, which was used for IMT. Modern semantic networks are usually implemented as semantic graphs, that are networks that represent semantic relationships between concepts where concepts are the vertices of the graph and edges represent semantic relations between them. Semantic networks have multiple uses. To date, machine translation is not among the most common ones.

A reason is that an interlingua representation in an open domain is difficult to achieve, and data-driven MT systems clearly outperform IMT for open-domain MT. Neural machine translation systems (NMT) are currently the state of the art for most language pairs (Bojar et al., 2017). Despite the success of this kind of architecture, it suffers from the same problem as other data-based translation systems: large amounts of parallel data must be available. To overcome this limitation, Artetxe et al. (2017) and Lample et al. (2017) introduce two unsupervised NMT methods that need only monolingual data but, up to now, they are far from the performance of seq2seq systems trained on bilingual corpora.

In this work, we investigate how a multilingual semantic network can be used for improving neural machine translation in general but specially for language pairs where not enough parallel data is available. We show how the inclusion of interlingual labels or synsets is beneficial in multilingual NMT (ML-NMT) systems and how they even allow beyond-zero-shot translation; that is, translation from languages that have not been seen in training. On the other hand, we explore a modern version of IMT, where the source text is codified into synsets and PoS tags and the translation into another natural language is learned by a seq2seq network.

Multilingual semantic networks have been used for machine translation mainly in statistical machine translation to deal with named entities and out-of-vocabulary words (Du et al., 2016; Srivastava et al., 2017). These issues are even more relevant in NMT because of the limited vocabulary that can be used to train the systems. However, the insights of seq2seq systems such as the difficulty to copy strings from the source into the target, make the integration a particular challenge.

The rest of the paper is organised as follows. Section 2. introduces BabelNet, the semantic network used for our experiments. Section 3. describes the NMT architecture and how the semantic information is included. Next, Section 4. describes the experiments and Section 5. analyses the results. Finally, Section 6. summarises and draws conclusions.

2. BabelNet

BabelNet (Navigli and Ponzetto, 2012) is a multilingual semantic network connecting concepts and named entities via *Babel synsets*. With 6 millions concepts and almost 8 millions named entities, the network covers 746 million word senses in 271 languages. This long list of languages, from Abkhazian to Zulu, includes many languages for which it is difficult to obtain parallel corpora.

Most of the concepts and named entities in BabelNet come from (Open Multilingual) WordNet, Wikipedia, Wikidata, Wiktionary and OmegaWiki. A synset groups these elements in different languages and treats them as synonyms in a language-independent way. The network also includes

	BabelNet				TED corpus	
Language (iso code)	Lemmas	Synsets	Senses	Synonym/Synset	Synsets	Coverage (%)
English (en)	11,769,205	6,667,855	17,265,977	2.59	28,445	27.25
French (fr)	5,301,989	4,141,338	7,145,031	1.73	–	–
German (de)	5,109,948	4,039,816	6,864,767	1.70	34,022	23.50
Spanish (es)	5,022,610	3,722,927	6,490,447	1.74	–	–
Dutch (nl)	4,416,028	3,817,696	6,456,175	1.69	27,720	26.25
Italian (it)	4,087,765	3,541,031	5,423,837	1.53	27,172	29.00
Romanian (ro)	3,009,318	2,697,720	3,384,256	1.25	24,375	27.25

Table 1: Statistics of BabelNet for the languages used in the experiments and coverage of the corpus with Babel synsets.

the lexico-semantic relations from WordNet and Wikipedia, but this information is not currently used in our approach, which focuses on the cross-language nature of synsets.

The left-hand side of Table 1 shows the key BabelNet figures for the seven languages used in our work. We observe a considerable gap between the number of lemmas covered in English and the remaining languages. However, as we show in Section 4.1., the difference does not translate into a significantly different coverage of the corpus. In what follows, languages are named by the ISO 329-1 code shown in the same table.

3. Seq2seq Neural Machine Translation

State-of-the-art NMT systems are seq2seq architectures with recurrent neural networks (RNN) (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2014). Briefly, an encoder projects source sentences into an embedding space and a decoder generates target sentences from the encoder embeddings.

Let $s = (x_1, \ldots, x_n)$ be a source sentence of length n. The encoder encodes s as a context vector at each word position, $\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n\}$, where each component is obtained by concatenating the forward ($\overrightarrow{\mathbf{h}}_i$) and backward ($\overleftarrow{\mathbf{h}}_i$) encoder RNN hidden states:

$$\mathbf{h}_i = \left[\overleftarrow{\mathbf{h}}_i, \overrightarrow{\mathbf{h}}_i\right] \quad (1)$$

with recurrent units

$$\overleftarrow{\mathbf{h}}_i = \tanh\left(\mathbf{W}_\mathbf{x} \|_{k=1}^{|F|} E_{xk} x_{ik} + \mathbf{U}_\mathbf{x} \overleftarrow{\mathbf{h}}_{i-1}\right) \quad (2)$$

$$\overrightarrow{\mathbf{h}}_i = \tanh\left(\mathbf{W}_\mathbf{x} \|_{k=1}^{|F|} E_{xk} x_{ik} + \mathbf{U}_\mathbf{x} \overrightarrow{\mathbf{h}}_{i-1}\right), \quad (3)$$

where $\mathbf{W}_x$ and $\mathbf{U}_x$ are trainable weight matrices, $\mathbf{E}_\mathbf{x}$ is the matrix of the source embeddings, and $\|$ is the concatenation operator. In the most simple case, the system is only trained with words so, $|F| = 1$, and $\mathbf{E}_\mathbf{x}$ corresponds to the matrix of word embeddings. Semantic information can be included as additional factors to the word representations. In this case, one considers two factors, $|F| = 2$, and concatenates synset embeddings to word embeddings which are learned independently. Other features and kinds of operations such as sum or multiplication could be used, the ones described here are those applied in our experiments.

Defined in this way, factors do not affect the decoding architecture. Let $t = (y_1, \ldots, y_m)$ be a target sentence of length m. The recurrent hidden state of the decoder $\mathbf{z}_j$ is computed using its previous hidden state $\mathbf{z}_{j-1}$, as well as the continuous representation of the previous target word $\mathbf{t}_{j-1}$ and the weighted context vector $\mathbf{q}_j$ at time step j:

$$\mathbf{z}_j = g(\mathbf{z}_{j-1}, \mathbf{t}_{j-1}, \mathbf{q}_j) \quad (4)$$

$$\mathbf{t}_{j-1} = \mathbf{E}_\mathbf{y} \cdot \mathbf{y}_{j-1}, \quad (5)$$

where g is a non-linear function and $\mathbf{E}_\mathbf{y}$ is the matrix of the target embeddings. The weighted context vector $\mathbf{q}_j$ is calculated by the *attention mechanism* as described in Bahdanau et al. (2014). Its function is to assign weights to the context vectors in order to selectively focus on different source words at different time steps of the translation and it is calculated as follows:

$$a(\mathbf{z}_{j-1}, \mathbf{h}_i) = \mathbf{v}_a \cdot \tanh(\mathbf{W}_a \cdot \mathbf{z}_{j-1} + \mathbf{U}_a \cdot \mathbf{h}_i) \quad (6)$$

$$\alpha_{ij} = \mathrm{softmax}\left(\mathrm{a}(\mathbf{z}_{\mathrm{j}-1}, \mathbf{h}_{\mathrm{i}})\right), \quad \mathbf{q}_j = \sum_i \alpha_{ij} \mathbf{h}_i \quad (7)$$

Finally, the probability of a target word is given by the following softmax activation (Sennrich et al., 2017):

$$p(y_j|\mathbf{y}_{<j}, \mathbf{x}) = p(y_j|\mathbf{z}_j, \mathbf{t}_{j-1}, \mathbf{q}_j) = \mathrm{softmax}\left(\mathbf{p}_j \mathbf{W}\right) \quad (8)$$

$$\mathbf{p}_j = \tanh\left(\mathbf{z}_j \mathbf{W}_{p1} + \mathbf{E}_\mathbf{y}[y_{j-1}] \mathbf{W}_{p2} + \mathbf{q}_j \mathbf{W}_{p3}\right) \quad (9)$$

where $\mathbf{W}_{p1}, \mathbf{W}_{p2}, \mathbf{W}_{p3}, \mathbf{W}$ are trainable matrices.

The number of target words in these systems is limited by the complexity of the training. The larger the vocabulary is, the higher the computing time and the memory needed. Usually, less than 100 k unique words are used.

4. Experimental Settings

4.1. Corpora

We use the *en-de-ro-it-nl* TED corpus provided for the IWSLT 2017 multilingual task (Cettolo et al., 2017). It includes 9161 talks in five languages, 4,380,258 parallel sentences when all the language pairs are considered. The intersection of talks among languages is high, 7945 documents are common to all of them, and therefore the same sentence is available in multiple languages. Notice that the size of the corpus is small as compared to standard collections of bilingual corpora —the WMT[1] en-fr set contains 36 M sentence pairs and the en-de one 5 M for instance. However, its multilingual nature makes it adequate for this study.

[1] http://statmt.org/wmt14/translation-task.html

SYSTEM w:	$< 2en >$ es war ein riesiger Erfolg
	$< 2en >$ è stato un enorme successo
SYSTEM wb:	$< 2en >$|- es|- war|- ein|- riesiger|- Erfolg|bn:15350982n
	$< 2en >$|- è|bn:00083181v stato|bn:00083181 un|- enorme|bn:00102268a successo|bn:00078365n
SYSTEM b:	PRONOUN VERB DETERMINER ADJECTIVE bn:15350982n
	bn:00083181v bn:00083181v DETERMINER bn:00102268a bn:00078365n

Figure 1: Example sentence of *tst2010* in German and Italian encoded to be translated into English for the three systems introduced in Section 4.2.: w, wb and b.

$< 2en >$|- es|- **war|bn:00083181** ein|- riesiger|- **Erfolg|bn:15350982n**
$< 2de >$|- and|- it|- **was|bn:00083181v** a|- huge|bn:00098905a **success|bn:00075023n**
$< 2en >$|- ed|- **è|bn:00083181v stato|bn:00083181v** un|- enorme|bn:00102268a **successo|bn:00078365n**
$< 2en >$|- en|- het|- **was|bn:00083181v** een|- groot|- **succes|bn:06512571n**
$< 2en >$|- și|bn:00012706n a|- **fost|bn:00083181v** un|- mare|bn:00098342a **succes|bn:00075024n**

Figure 2: Sentence extracted from *tst2010* in the five languages of the TED corpus en-de-ro-it-nl. The encoding as input to system wb shows differences and similarities of Babel synsets among languages.

We annotate the documents with a coarse-grained part of speech tagset (PoS), lemma and Babel synsets. Our PoS tag set consists of 10 elements defined to be compatible with the one in the BabelNet ontology {NOUN, VERB, PREPOSITION, PRONOUN, DETERMINER, ADVERB, ADJECTIVE, CONJUNCTION, ARTICLE, INTERJECTION}. The IXA pipeline (Agerri et al., 2014) is used to annotate en, de, es and fr documents with PoS and TreeTagger (Schmid, 1994) for nl, ro and it. The original tags are then mapped to our common reduced tagset[2]. The same tools are used to annotate the texts with lemmas.

Only a subset of PoS tags is enriched with their synset information. We select (*i*) nouns —including named entities, foreign words and numerals, (*ii*) adjectives, (*iii*) adverbs and (*iv*) verbs. In addition, we explicitly mark negation particles with a tag NEG and include them here to account for their semantics. Since 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, we select the BabelNet ID as the ID according to the BabelNet ordering of IDs: "*(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*" (Navigli, 2013, p. 35).

With this procedure, 26.5% of the corpus is covered by synset identifiers and the remaining 73.5% only by PoS tags, where the coverage per language is similar and ranges from 23.5% to 29.0%, see Table 1.

4.2. NMT Systems

Our systems are NMT engines trained with Nematus (Sennrich et al., 2017). We train three systems:

w: A many-to-many NMT engine trained on parallel corpora for the several language pairs simultaneously. As in Johnson et al. (2017) and similarly to Ha et al. (2016), the engine is trained with the only addition of a tag in the source sentence to account for the target language "<2trg>". We only consider those sentences with less than 50 tokens for training, that is 2,113,917 parallel sentences (39,393,037 tokens).

wb: A many-to-many factored NMT engine (Sennrich and Haddow, 2016) trained on the same corpus as before but enriched with Babel synsets as an additional factor.

b: A one-to-one NMT system trained on the part of the corpus with English as target. All the source languages are encoded as Babel synsets instead of words; for any word without a known synset, we use the PoS. This way, we obtain 868,226 parallel sentences (15,684,750 tokens).

Figure 1 shows example sentences coded according to each system.

Regarding the system's 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 embeddings is always 506; for the factored system wb we reserve 300 dimensions for words and 206 for synsets. All the systems have a maximum common vocabulary of 150 k. Systems w and wb add 2 k for subword units segmented using Byte Pair Encoding (BPE) (Sennrich et al., 2016). Subwords in the source sentence are annotated with the same factors as the corresponding complete word. There is no BPE segmentation in system b. For decoding, we use an ensemble with the last four models at intervals of 10000 mini-batches and a beam size of 10.

5. Results and Discussion

Table 2 shows the translation performance of the three systems defined in the previous section on the 2010 IWSLT test set (*tst2010*), a test set build up with unseen TED talks. Systems are trained on en, de, ro, it and nl data (top rows); fr and es (bottom rows) have not been seen in training and correspond to what we call beyond-zero-shot translation. Boldfaced scores in the table mark the best system for a

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

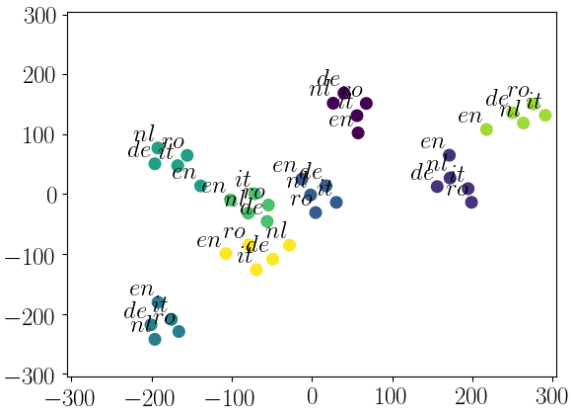

Figure 3: 2D t-SNE representation of the context vectors of the first 8 source sentences of *tst2010* for system wb. The same sentence has the same colour in different languages.

language pair and, when systems are not statistically significantly different from the best one with at least a p-value of 0.01, we mark them as well. Bootstrap resampling is used to estimate statistical significance (Koehn, 2004).

For the languages with training data, we observe that the addition of cross-language synsets as factors moderately improves the translation quality as measured by BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014). Both are lexical metrics that essentially count the number of n-gram matches between a system translation and the reference translation(s). We use the version of METEOR that considers matches between words, stems, synonyms and paraphrases.

Cross-language IDs are of special interest to ML-NMT systems because they can help to cluster together sentences according to their meaning and irrespective of their language. Since a word vector in our factored NMT (wb) has the top elements representing the word itself and the bottom ones representing the synset ID, different words with same synset share a common part of the representation. In fact, such a clustering is already done in ML-NMT systems, but the quality is worse the more distant languages are (España-Bonet et al., 2017) and synsets can help to overcome this distance. In our case, by using the ML-TED corpus, we train the systems with the same sentence in different languages, so this grouping is already eased by construction, as becomes apparent through a graphical representation of the sentences. Figure 3 depicts a 2D t-SNE representation (Van Der Maaten, 2014) of the context vectors of 8 sentences of the test set with the wb system. The clustering by sentence (colour) is evident in the plot but we obtain very similar clustering visualisations with the w system. Since the initial grouping is already good, the addition of the synsets improves the translation by only 0.23 points on average.

The quality of the synset annotation is also relevant for performance. One of the major issues in our setting is the fact that the top synset in a language does not always correspond to the top synset in another one. The example sentence in Figure 2 is an extreme case where the word *success* has five different IDs depending on the language. The verb *to be* on the contrary, is identified as *bn:00083181v* in all of them.

	BLEU			METEOR		
	w	wb	b	w	wb	b
$de2en$	**32.6**	**33.0**	17.5	33.1	**33.5**	24.2
$it2en$	**33.5**	**33.2**	21.4	**33.9**	**34.0**	27.4
$nl2en$	36.2	**36.6**	15.0	34.7	**34.9**	21.5
$ro2en$	**34.3**	**34.8**	19.6	**34.4**	**34.6**	25.9
$fr2en$	2.4	5.1	**7.3**	11.2	16.7	**17.5**
$es2en$	3.1	6.7	**11.3**	12.0	18.4	**20.7**

Table 2: Automatic evaluation of the systems defined in Section 4.2. on *tst2010*. Boldfaced scores indicate the best systems; systems not statistically significantly different from the best one ($p = 0.01$) are also boldfaced.

Improving the cross-linguality in the synset annotation is a key aspect to achieve further improvements. Besides, as stated in Section 4.1., we did not perform any word sense disambiguation for retrieving the synset but took the top ID, so we are missing relevant information for translation which could also help to gather the truly interlingua IDs.

Even with these identified limitations, the factored system wb already improves on the word system w and this is even more evident in the case of languages that have not been seen at training time. The last two rows in Table 2 display the results when translating from unseen es and fr into en. In this case, the system does not have the vocabulary of the language, so a BLEU score of 2.4 ($fr2en$) and 3.1 ($es2en$) is obtained mainly thanks to identical named entities, digits and cognates between the languages. The inclusion of synsets is in this case more important, because words sharing the synset ID can be now translated and that increases the BLEU scores to 5.1 ($fr2en$) and 6.7 ($es2en$), +3.2 BLEU points in average. Similar differences are seen with METEOR. Still the numbers are far from those obtained for languages seen in training.

System b is totally different. Here the source words are not used at all and we keep *what is left after removing the structural particularities of the base language* as Richens (1956) suggested to encode a source sentence. For a language pair with parallel corpora this representation is clearly worse than the original one because all the morphological information and even the semantics of prepositions, determiners and conjunctions is lost. However, the semantics of content words is kept in an interlingual way and that improves the translation of unseen languages, +6.5 BLEU and +7.5 METEOR points on average as shown in Table 2.

Comparing b with similar systems trained on monolingual data, we observe that the translation is possible because we use multiple languages on the source side, and the network learns different combinations to encode the same expression. For instance, both *"PRONOUN VERB DETERMINER ADJECTIVE bn:15350982n"* and *"bn:00083181v bn:00083181v DETERMINER bn:00102268a bn:00078365n"* should be translated as *"it was a huge success"* (Figure 1). This diversity is important to accommodate new languages.

Strengths and weaknesses of the three systems can be seen in the example translation shown in Figure 4. For the languages with training data, w and wb provide the same

	SYSTEM w:	SYSTEM wb:	SYSTEM b:
$de2en$	and it was a huge success	and it was a huge success	and it 's a huge success
$it2en$	and it was a huge success	and it was a huge success	and it was a huge success
$nl2en$	and it was a big success	and it was a big success	and this is a huge success
$ro2en$	and it was a great success	and it was a great success	and it was a great success
$fr2en$	it 's the facade of a great success	and the Khan has been a great winner	but there was a big winner
$es2en$	y is a great deal	y is a great marker	but it 's a great winner

Figure 4: Example sentence of *tst2010* in the languages of the study translated into English by the three systems introduced in Section 4.2.

translation for this simple sentence. For fr and es, where there is no training data, some of the words are cognates and have been seen in other languages (*gran*/es, *grand*/fr, *succes*/fr; *gran*/it, *grand*/en, *succes*/ro) while some others have not (*fue*/es, *ça*/fr, *été*/fr). In the latter case the w system just builds the translation as the concatenation of seen BPE subunits (*ça a été*/$fr \Rightarrow$ *it's the facade of*/en), while the wb system is able to recognise the verb thanks to the synset (*été*|*bn:00083181v* $\Rightarrow$ *has been*). As before, b behaves differently. When the synset is correctly assigned, the system can translate the adjective (*huge*, *big*, *great*) even if the ID differs for each source language. As shown in Figure 1, *riesiger* in the German sentence could not be mapped to a synset, so system b translates it from the source token *ADJECTIVE*. In this particular case the translation is correct because during training the system has learnt that *huge* is the most probable translation for *ADJECTIVE* when it goes before *Erfolg*. However, part-of-speech tags cannot always be translated properly and we obtain different choices for *CONJUNCTION* (*and*, *but*) and *PRONOUN* (*it*, *this*, *there*) depending on the sentence. Conjugations might not be correctly translated either: *VERB* (*'s*, *is*, *was*).

The previous example shows how in order to make the most of this architecture, one would need an additional abstraction step for non-content words and making morphology explicit in the source side, and then in the corresponding generation step in the decoder side. That would even allow to train a synset2target NMT system using only monolingual data. These refinements are left as future work.

6. Summary and Conclusions

We have shown two different ways to include the knowledge encoded in semantic networks in NMT systems. The first one, system wb, adds interlingual Babel synsets as a factor. This way, we obtain moderate improvements in ML-NMT translation for known languages, and more than 3 BLEU points for languages not seen in training. The second one, b, encodes the input as a sequence of Babel synsets completed with PoS tags and entirely ignoring the specific words of the source language. This way, we further improve translation for languages not seen in training (beyond zero shot) by more than 6.5 BLEU and 7.5 METEOR points on average.

The next natural step is to design these systems so that they can be trained on monolingual corpora only. To do this, we need first to better choose (i.e. properly disambiguate) the synset of a word so that it is the same irrespective of the language. Second, one needs to add abstraction and generation layers to deal with morphology and non-content words in the target language.

Notice that the methodology used benefits from the ability of seq2seq models to learn in multilingual settings, so it is not exclusive to NMT and it can be also applied to multilingual/crosslingual neural text summarisation or question answering systems for example.

7. Acknowledgements

The project on which this paper is based was partially funded by the German Federal Ministry of Education and Research under the funding code 01IW17001 (Deeplee) and by the Leibniz Gemeinschaft via the SAW-2016-ZPID-2 project (CLuBS). Responsibility for the content of this publication is with the authors. The authors are grateful to R. Navigli for the useful discussions and comments.

8. Bibliographical References

Agerri, R., Bermudez, J., and Rigau, G. (2014). 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, may. European Language Resources Association (ELRA).

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2017). Unsupervised neural machine translation. *CoRR*, abs/1710.11041.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*, abs/1409.0473, September.

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

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

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–

1734, Doha, Qatar, October. Association for Computational Linguistics.

Denkowski, M. and Lavie, A. (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Du, J., Way, A., and Zydron, A. (2016). Using babelnet to improve oov coverage in smt. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

España-Bonet, C., Varga, A. C., Barrón-Cedeño, A., and van Genabith, J. (2017). An Empirical Analysis of NMT-Derived Interlingual Embeddings and their Use in Parallel Sentence Identification. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1340–1350, December.

Ha, T., Niehues, J., and Waibel, A. H. (2016). Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. In *Proceedings of the International Workshop on Spoken Language Translation*, Seattle, WA, November.

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

Koehn, P. (2004). Statistical Significance Tests for Machine Translation Evaluation. In Dekang Lin et al., editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Lample, G., Denoyer, L., and Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Navigli, R. (2013). A quick tour of babelnet 1.1. In *CICLing (1)*, volume 7816 of *Lecture Notes in Computer Science*, pages 25–37. Springer.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Association of Computational Linguistics*, pages 311–318.

Richens, R. H. (1956). Preprogramming for mechanical translation. *Mechanical Translation*, 3(1):20–25, July.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Sennrich, R. and Haddow, B. (2016). Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 83–91, Berlin, Germany, August. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). 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*.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Miceli Barone, A. V., Mokry, J., and Nadejde, M. (2017). 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*, pages 65–68, Valencia, Spain, April. Association for Computational Linguistics.

Srivastava, A., Rehm, G., and Sasaki, F. (2017). Improving machine translation through linked data. *The Prague Bulletin of Mathematical Linguistics (PBML)*, 108:355–366, 6.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Van Der Maaten, L. (2014). Accelerating t-SNE Using Tree-based Algorithms. *Journal of Machine Learning Research*, 15(1):3221–3245, January.