doi:10.1017/S1351324918000293

A neural approach for inducing multilingual resources and natural language processing tools for low-resource languages

O. ZENNAKI[1,2], N. SEMMAR[1] and L. BESACIER[2]

[1]*CEA, LIST, Vision and Content Engineering Laboratory, Gif-sur-Yvette, France*

[2]*Laboratory of Informatics of Grenoble, Univ. Grenoble-Alpes, Grenoble, France*

e-mails: othman.zennaki@cea.fr, nasredine.semmar@cea.fr, laurent.besacier@imag.fr

(*Received 6 March 2017; revised 4 June 2018; accepted 5 June 2018*)

Abstract

This work focuses on the rapid development of linguistic annotation tools for low-resource languages (languages that have no labeled training data). We experiment with several cross-lingual annotation projection methods using recurrent neural networks (RNN) models. The distinctive feature of our approach is that our multilingual word representation requires only a parallel corpus between source and target languages. More precisely, our approach has the following characteristics: (a) it does not use word alignment information, (b) it does not assume any knowledge about target languages (one requirement is that the two languages (source and target) are not too syntactically divergent), which makes it applicable to a wide range of low-resource languages, (c) it provides authentic multilingual taggers (one tagger for N languages). We investigate both uni and bidirectional RNN models and propose a method to include external information (for instance, low-level information from part-of-speech tags) in the RNN to train higher level taggers (for instance, Super Sense taggers). We demonstrate the validity and genericity of our model by using parallel corpora (obtained by manual or automatic translation). Our experiments are conducted to induce cross-lingual part-of-speech and Super Sense taggers. We also use our approach in a weakly supervised context, and it shows an excellent potential for very low-resource settings (less than 1k training utterances).

1 Introduction

Annotating linguistic resources consists of adding information of interpretative nature to the original raw data (Garside, Leech and McEnery 1997). This information may be terminological, lexical, morphological, syntactic or semantic. Linguistic resources can be lexicons, transcriptions of dialogues, text corpora, etc. (Veronis 2000).

Annotating corpora with linguistic information (part-of-speech tagging, sense tagging, syntactic analysis, named entity identification and semantic role annotation) involves significant human efforts. The availability of parallel corpora has recently led to several research studies on cross-lingual annotation projection. The idea is to explore the use of unsupervised (no-supervision on the targets languages) approaches which project labels from the (resource-rich) *source* language (such as English which has large amounts of annotated corpora and several analysis tools available) to the (under-resourced) *target* language with less resources. The goal of cross-language projection is, on the one hand, to provide all languages with linguistic annotations, and on the other hand, to automatically induce text analysis tools for these languages.

In this work, we address the task of automatic building of multilingual linguistic resources and tools. This article presents the use of unsupervised approaches to learn multilingual Natural Language Processing (NLP) tools from parallel corpora with only annotations in the source language.

Contributions. We investigate different architectures of RNNs — unidirectional RNN (URNN) and bidirectional RNN (BRNN) — for multilingual sequence labeling tasks. Two NLP tasks are considered: Part-of-speech (POS) tagging and Super Sense tagging (SST) (Ciaramita and Altun 2006). Our RNN architectures demonstrate very good results on unsupervised training for new target languages. In addition, we show that the integration of POS information in RNN models is useful to build a multilingual coarse-grain semantic (Super Sense) tagger. For this, a simple and efficient way to take into account low-level linguistic information for more complex sequence labeling tasks is proposed. Semi-supervised scenarios (where a small amount of annotated target data is available) are also investigated in this journal article.

Outline. The remainder of the article is organized as follows. Section 2 reviews related work. Section 3 describes our cross-language annotations projection approaches based on RNN. Section 4 (POS tagging) and Section 5 (SST) present our empirical studies and experimental results. We finally conclude the article in Section 6 and present our future research directions.

2 Related work

We present below most approaches for cross-lingual projection based on word alignment or representation learning.

2.1 Cross-lingual projection based on word alignment

The availability of parallel corpora has recently led to several research works exploring the use of unsupervised or semi-supervised cross-lingual annotation projection based on word alignment information. Through word alignments in parallel corpora, the annotations are transferred from the (resource-rich) *source* language to the (potentially under-resourced) *target* language.

2.1.1 Unsupervised methods

Cross-lingual projection of linguistic annotations was pioneered by Yarowsky, Ngai and Wicentowski (2001) who proposed to transfer annotations from resource-rich languages onto resource-poor languages, based only on word alignments from a parallel corpus. After alignment, the source language is annotated, and annotations are projected to the target language. The resulting (noisy) annotations are used in conjunction with robust learning algorithms to build cheap unsupervised NLP tools. This approach has been successfully used to transfer several linguistic annotations between languages. Examples include POS (Das and Petrov 2011; Duong, Cook, Bird and Pecina 2013), named entities (Kim, Toutanova and Yu 2012), syntactic constituents (Jiang, Liu and Lü 2011), word senses (Bentivogli, Forner and Pianta 2004; van der Plas and Apidianaki 2014) and semantic role labels (Pado and Pitel 2007; Annesi and Basili 2010).

In recent years, several works have investigated transfer learning (Pan and Yang 2010) in NLP. We can cite the work of Jiang *et al.* (2015) who applied transfer learning on annotation adaptation. They implemented several models for transferring annotations of a source corpus to the annotation format of another target corpus. These models are based on a transfer classifier which learns correspondence regularities between annotation guidelines from a parallel annotated corpus, which has two kinds of annotations for the same data. In this specific application, the source corpus and the target corpus are in the same language. In the same vein, Passban, Liu and Way (2017) implemented statistical and neural machine translation engines that are trained on one language pair but are used to translate another language. They trained a reliable model for a high-resource language, and then they exploited cross-lingual similarities in order to adapt the model to work for a close language with few resources.

In order to build our baseline system, we use direct transfer which is similar to the method described in Yarowsky *et al.* (2001). First, we tag the source side of the parallel corpus using an available supervised tagger. Next, we align words in the parallel data using GIZA++ (Och and Hermann 2000) to find out corresponding source and target words. Then, we project tags in the target language using following projection criteria:

- For 1-to-1 alignments, we project tags directly.
- For N-to-1 mappings, we project the tag of the word at the position $N/2$ rounded up to the next whole number.
- The unaligned words (target) are tagged with their most frequent associated tag in the corpus.

We finally train a tagger on the target language using *Trigrams'n'Tags* (TnT) statistical tagger (Brants 2000).

2.1.2 Semi-supervised methods

Cross-lingual projection requires a parallel corpus and word alignment between source and target languages. Many automatic word alignment tools are available, such as GIZA++ which implements IBM models (Brown *et al.* 1993). However, the noisy (non-perfect) outputs of these methods are a serious limitation for the annotation projection based on word alignments (Fraser and Marcu 2007). To deal with these limitations, recent studies have proposed to combine projected labels with partially supervised monolingual information in order to filter out invalid label sequences. For example, Li, Graça and Taskar (2012), Täckström *et al.* (2013) and

Wisniewski *et al.* (2014) have proposed to improve projection performance by using a dictionary of valid tags for each word (coming from Wiktionary[1]).

2.2 Projection based on cross-lingual representation learning

In another vein, recent studies based on cross-lingual representation learning methods have been proposed to avoid using such pre-processed and noisy alignments for label projection. First, these approaches learn language-independent features, across many different languages (Durrett, Pauls and Klein 2012; Al-Rfou, Perozzi and Skiena 2013; Täckström, McDonald and Nivre 2013; Luong, Pham and Manning 2015; Gouws and Søgaard 2015; Gouws, Bengio and Corrado 2015). Then, the induced representation space is used to train NLP tools by exploiting labeled data from the source language and applying them in the target language. Cross-lingual representation learning approaches have achieved good results in different NLP applications such as cross-language SST and POS tagging (Gouws and Søgaard 2015), cross-language named entity recognition (Täckström, McDonald and Uszkoreit 2012), cross-lingual document classification and lexical translation (Gouws *et al.* 2015), cross language dependency parsing (Durrett *et al.* 2012; Täckström *et al.* 2013) and cross-language semantic role labeling (Titov and Klementiev 2012).

Our approach described in the next section, is inspired by these works since we also try to induce a common language-independent feature space (cross-lingual words embeddings). Unlike Durrett *et al.* (2012) and Gouws and Søgaard (2015), who use bilingual lexicons, and unlike Luong *et al.* (2015) who use word alignments between source and target languages,[2] our common multilingual representation is very agnostic. We use a simple (multilingual) vector representation based on the occurrences of source and target words in a parallel corpus and we let the RNN learn the best internal representations (corresponding to the hidden layers) specific to the task (SST, POS tagging or other tasks).

In this work, we learn a cross-lingual POS tagger (multilingual POS tagger if a multilingual parallel corpus is used) based on a recurrent neural network (RNN) on the source labeled text and apply it to tag target language text. We explore URNN and BRNN architectures, respectively. We also show that the proposed architecture is well suited for lightly supervised training (adaptation). Starting from the intuition that low-level linguistic information is useful to learn more complex taggers, we also introduce three new RNN variants to take into account external (POS) information in a more complex task (multilingual SST).

3 Our approach: Using recurrent neural networks in cross-lingual projection of annotations

To avoid projecting label information from deterministic and error-prone word alignments, we propose to represent the bilingual word information intrinsically in

[1] http://www.wiktionary.org/

[2] To train a bilingual representation regardless of the task.

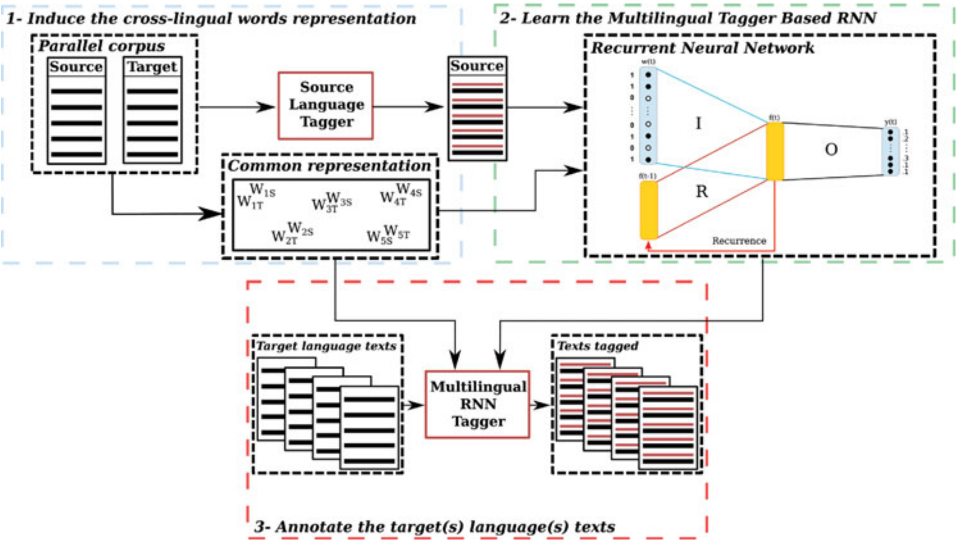

Fig. 1. (Colour online) Overview of the proposed model architecture for inducing multilingual RNN taggers.

an RNN architecture. Then, this RNN is used as a multilingual sequence labeling model (we investigate POS tagging and SST tasks).

3.1 Proposed method

We propose a method for learning multilingual sequence labeling tools based on RNN, as can be seen in Figure 1. In our approach, a parallel or multi-parallel corpus between a resource-rich language and one or many target (potentially under-resourced) language(s) is used to extract common (multilingual) and agnostic words representation. These representations, which rely on sentence-level alignment only, are used with the source side of the parallel/multi-parallel corpus to learn a neural network tagger in the source language. Since a common representation of source and target words is chosen, this neural network tagger is authentically multilingual and can also be used to tag texts in target language(s).

In our *agnostic* representation, we associate to each word (in source *and* target vocabularies) a common vector representation, namely V_w (equation (1)).

$$V_w = \begin{cases} V_{wi} = 1 \;\; if \;\; w \in S_i \\ V_{wi} = 0 \, else \end{cases} \tag{1}$$

S_i, $i = 1, \ldots, N$ is the ith bisentence of the parallel corpus, where N is the number of parallel sentences (bisentences in the parallel corpus).

The idea is that, in general, a source word and its target translation appear together in the same bisentences and their vector representations are close. We can then use the RNN tagger, initially trained on source side, to tag the target side (because of our *common vector representation*). This simple representation does not require multilingual word alignments since the RNN learns the optimal internal

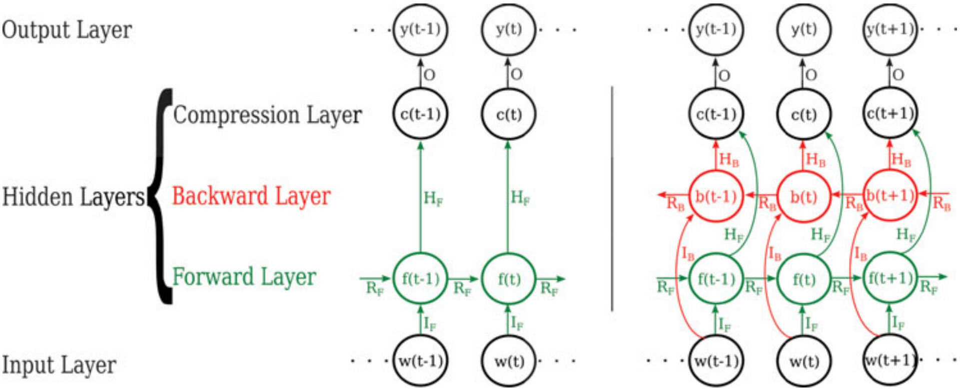

Fig. 2. (Colour online) High-level schema of RNN used in our work unidirectional (left schema) and bidirectional RNN (right schema).

representation needed for the annotation task (for instance, the hidden layers of the RNN can be considered as multilingual word embeddings).

3.2 Recurrent neural networks

There are two major architectures of neural networks: Feedforward (Bengio *et al.* 2003) and RNN (Schmidhuber 1992; Mikolov *et al.* 2010). Sundermeyer *et al.* (2013) showed that language models based on recurrent architecture achieve better performance than language models based on feedforward architecture. This is due to the fact that RNNs do not use a context of limited size. This property led us to use, in our experiments, the Elman recurrent architecture (Elman 1990), in which recurrent connections occur at the hidden layer level.

We consider in this work two Elman RNN architectures (see Figure 2): URNN and BRNN. In addition, to be able to include low-level linguistic information in our architecture designed for more complex sequence labeling tasks, we propose three new RNN variants to take into account external information (such as POS tags) for a more complex task: multilingual SST.

3.2.1 Unidirectional RNN

In the Elman URNN, the recurrent connection is a loop at the hidden layer level. This connection allows URNN to use at the current time step hidden layer's states of previous time steps. In other words, the hidden layer of URNN represents all previous history and not just $n-1$ previous inputs, thus the model can theoretically represent long context.

The architecture of the URNN considered in this work is shown in Figure 2. In this architecture, we have four layers: input layer, forward (also called recurrent or context layer), compression hidden layer and output layer. All neurons of the input layer are connected to every neuron of forward layer by weight matrix I_F and R_F, the weight matrix H_F connects all neurons of the forward layer to every neuron of compression layer and all neurons of the compression layer are connected to every neuron of output layer by weight matrix O.

The input layer consists of a vector $w(t)$ that represents the current word w_t in our common words representation (all input neurons corresponding to current word w_t are set to 0 except those that correspond to bisentences containing w_t, which are set to 1), and of vector $f(t-1)$ that represents output values in the forward layer from the previous time step. We name $f(t)$ and $c(t)$ the current time step hidden layers (our preliminary experiments have shown better performance using these two hidden layers instead of one hidden layer), with variable sizes (usually 80–1,024 neurons) and sigmoid activation function. These hidden layers represent our common language-independent feature space and inherently capture word alignment information. The output layer $y(t)$, given the input $w(t)$ and $f(t-1)$ is computed with the following steps:

$$f(t) = \Sigma(w(t).I_F(t) + f(t-1).R_F(t)) \tag{2}$$

$$c(t) = \Sigma(f(t).H_F(t)) \tag{3}$$

$$y(t) = \Gamma(c(t).O(t)) \tag{4}$$

Σ and Γ are the sigmoid and the softmax functions, respectively. The softmax activation function is used to normalize the values of output neurons to sum up to 1. After the network is trained, the output $y(t)$ is a vector representing a probability distribution over the set of tags. The current word w_t (in input) is tagged with the most probable output tag.

For many sequence labeling tasks, it is beneficial to have access to future in addition to the past context. So, it can be argued that our URNN is not optimal for sequence labeling, since the network ignores future context and tries to optimize the output prediction given the previous context only. This URNN is thus penalized compared with our baseline projection based on TnT (Brants 2000) which considers both left and right contexts. To overcome the limitations of URNN, a simple extension of the URNN architecture — namely, BRNN (Schuster and Paliwal 1997) — is used to ensure that context at previous and future time steps will be considered.

3.2.2 Bidirectional RNN

An unfolded BRNN architecture is given in Figure 2. The basic idea of BRNN is to present each training sequence forwards and backwards to two separate recurrent hidden layers (forward and backward hidden layers) and then somehow merge the results. This structure provides the compression and the output layers with complete past and future context for every point in the input sequence. Note that without the backward layer, this structure simplifies to a URNN.

The inference procedure for the unfolded bidirectional network for one time slice $1 \leq t \leq T$ to determine all predicted outputs, was described in the reference paper on BRNN (Schuster and Paliwal 1997), and can be summarized as follows:

- Compute the backward hidden layers states from $t = T$ to $t = 1$ (equation (5)):

$$b(t) = \Sigma(w(t).I_B(t) + b(t+1).R_B(t)) \tag{5}$$

- Use the backward hidden layers states to compute the compression layers c(t) (equation (6)) and output layers $y(t)$ (equation (4)) from $t = 1$ to $t = T$:

$$c(t) = \Sigma(f(t).H_F(t) + b(t).R_b(t)) \tag{6}$$

3.2.3 Network training

The first step in our approach is to train the neural network, given a parallel corpus (training corpus) and a validation corpus (different from train data) in the source language. In typical applications, the source language is a resource-rich language (which already has an efficient tagger or manually tagged resources). Our RNN models are trained by stochastic gradient descent using usual back-propagation and back-propagation through time algorithms (Rumelhart, Hinton and Williams 1985). We learn our RNN models with an iterative process on the tagged source side of the parallel corpus. After each epoch (iteration) in training, validation data is used to compute per-token accuracy of the model. After that, if the per-token accuracy increases, training continues in the new epoch. Otherwise, the learning rate is halved at the start of the new epoch. Eventually, if the per-token accuracy does not increase anymore, training is stopped to prevent over-fitting. Generally, convergence takes 5–10 epochs, starting with a learning rate $\alpha = 0.1$.

The second step consists in using the trained model as a target language tagger (using our common vector representation). It is important to note that if we train on a multilingual parallel corpus with N languages ($N > 2$), the same trained model will be able to tag all the N languages.

Hence, our approach assumes that the word order in both source and target languages are similar. In some languages such as English and French, word order for contexts containing nouns could be reversed most of the time. For example, the compound word *the European Commission* would be translated into *la Commission Européenne*. In order to deal with the word order constraints, we also combine the RNN model with the baseline cross-lingual projection model in our experiments. In case of syntactic divergence between source and target languages, another idea would be to pre-process text in target language to better match the text in source language. This could be done using linguistically motivated rules automatically extracted from typological databases such as the World Atlas of Language Structures[3]. This latter possibility is part of our future work.

3.3 Dealing with out-of-vocabulary words

For the words unseen in the initial parallel corpus (OOV words), their vector representation is a vector of zero values. Consequently, during testing, the RNN

[3] `http://wals.info`

model will use only the context information to tag the OOV words found in the test corpus. To deal with these types of OOV words,[4] we use the CBOW model of Mikolov *et al.* (2013) to replace each OOV word by its closest known word in the current OOV word context (to achieve this, we compute cosine similarity between word embeddings). Once the closest word is found, its common vector representation is used (instead of the vector of zero values) at the input of the RNN.

3.4 Combining simple cross-lingual projection and RNN models

Since the direct transfer model *M1* and RNN model *M2* use different strategies for tagging (TnT is based on Markov models, while RNN is a neural network), we assume that these two models can be complementary. To keep the benefits of each approach, we explore how to combine them with linear interpolation. Formally, the probability to tag a given word w is computed as

$$P_{M12}(t|w) = (\mu P_{M1}(t|w, C_{M1}) + (1 - \mu)P_{M2}(t|w, C_{M2})) \tag{7}$$

where C_{M1} and C_{M2} are the context of w considered by *M1* and *M2*, respectively. The relative importance of each model is adjusted through the interpolation parameter μ. The word w is tagged with the most probable tag, using function f described as

$$f(w) = \arg\max_{t}(P_{M12}(t|w)) \tag{8}$$

4 Multilingual part-of-speech tagging-based RNN

We applied our method to build RNN POS taggers for four target languages — French, German, Greek and Spanish — with English as the source language. No annotation is available for the target languages (we suppose these languages are under-resourced).

In order to determine the effectiveness of our common words representation described in Section 3.1, we also investigated the use of state-of-the-art bilingual word embeddings (using MultiVec Toolkit[5] (Bérard *et al.* 2016)) as input to our RNN tagger.

4.1 Light supervision (adaptation) of RNN model

While the unsupervised RNN (no-supervision on the target languages) model described in the previous section has not seen any annotated data in the target language, we also consider the use of a small amount of adaptation data (manually annotated in target language), relatively to the huge number of sentences used to learn supervised tagger, in order to capture target language specificity. In that case, the RNN model is adapted in a light supervision manner, using a small monolingual target corpus (manually annotated — the exact size used for adaptation is provided

[4] Words which do not have a known vector representation.
[5] `https://github.com/eske/multivec`

in Table 5 –) and the common vector representation of words (extracted from the initial parallel corpus). We plan to experiment with a gradually increasing amount of annotated data in the target language (from 100 to 10, 000 utterances). Such an approach is particularly suited for an iterative scenario where a user would post-edit (correct) the unsupervised POS tagger output in order to produce rapidly adaptation data in the target language (light supervision).

4.2 Data

For French as a target language, we used a train set of 10, 000 parallel sentences, a validation set of 1, 000 English sentences and a test set of 1, 000 French sentences, all extracted from the ARCADE II English–French corpus (Veronis *et al.* 2008). The test set is tagged with the French *TreeTagger* (Schmid 1995) and then manually checked.

For German, Greek and Spanish as a target language, we used training and validation data extracted from the Europarl corpus (Koehn 2005) which are a subset of the training data used in Das and Petrov (2011) and Duong *et al.* (2013). This choice allows us to compare our results with those of Das and Petrov (2011), Duong *et al.* (2013) and Gouws and Søgaard (2015). The train data set contains 65, 000 bisentences, a validation set of 10, 000 bisentences is also available. For testing, we use the same test corpora as (Das and Petrov 2011), (Duong *et al.* 2013) and (Gouws and Søgaard 2015) (bisentences from CoNLL shared tasks on dependency parsing (Buchholz and Marsi 2006)). The evaluation metric (*per-token* accuracy) and the *universal tagset* defined by Petrov, Das and McDonald (2012) (twelve tags common for most languages) are used for evaluation. This universal POS tagset defines the following twelve POS tags, which exist in similar form in most languages: NOUN (nouns), VERB (verbs), ADJ (adjectives), ADV (adverbs), PRON (pronouns), DET (determiners and articles), ADP (prepositions and postpositions), NUM (numerals), CONJ (conjunctions), PRT (particles), '.' (punctuation marks) and X (a catch-all for other categories such as abbreviations or foreign words).

For training, the English (source) sides of the training corpora (ARCADE II and Europarl) and of the validation corpora are tagged with the English *TreeTagger* toolkit. Using the matching provided by Petrov *et al.* (2012), we map TreeTagger tagset as well as tagset found in CoNLL shared task data to the common universal tagset.

In order to build our baseline unsupervised tagger (based on a Direct Cross-lingual Projection — see Section 2.2.1), we also tag the tagger side of the training corpus, with tags projected from English side through word-alignments established by GIZA++ (alignment models were trained using IBM Models 1–5). After the tags projection, a target language POS tagger based on TnT approach (Brants 2000) is trained.

The combined model is built for each considered language using cross-validation on the test corpus. First, the test corpus is split into two equal parts and on each part, we estimate the interpolation parameter μ (equation (7)) which maximizes the *per-token* accuracy score. Then each part of the test corpus is tagged using the

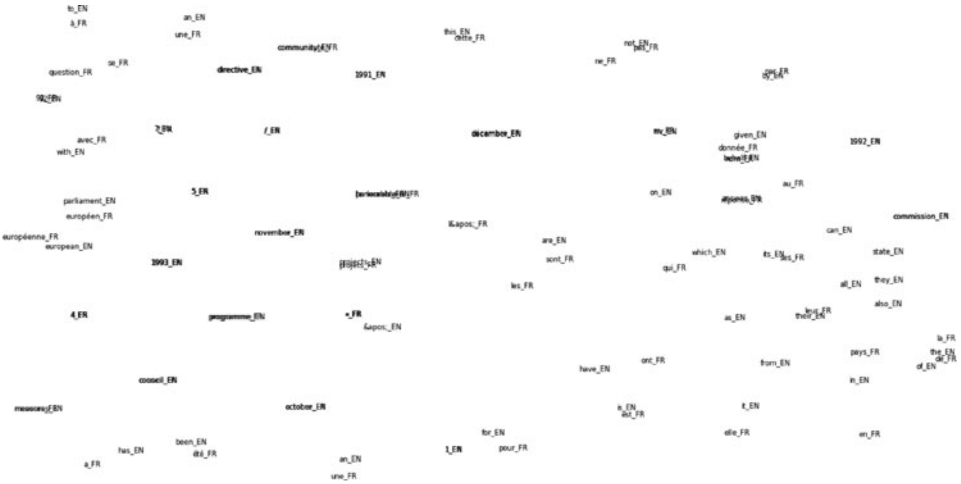

Fig. 3. A t-SNE visualization of RNN vector (word embeddings) representations of the same few frequent English and French words.

combined model tuned on the other part, and vice versa (standard cross-validation procedure).

We trained MultiVec bilingual word embeddings on the parallel Europarl corpus between English and each of the target languages considered.

4.3 Qualitative evaluation

As we have previously stated, we propose to represent the bilingual word information intrinsically in our RNN model by letting the RNN learn its internal words representation (bilingual word embeddings) using our common words representation (Section 3.1). We present here a qualitative evaluation of English–French bilingual word embeddings learned from the ARCADE II English–French corpus. We jointly visualized the RNN bilingual words representation of the most frequent words in English and French using the t-SNE (Van der Maaten and Hinton 2008) visualization tool. The embeddings are shown in Figure 3. We observe that the visualization of the embeddings of English words and their French translations are close and even in some cases they are overlapped. Therefore, it seems that our RNN model is able to represent bilingual word information.

4.4 Results and discussion

4.4.1 Results

Table 1 reports the results obtained for the unsupervised POS tagging. We note that the POS tagger based on BRNN has better performance than URNN, which means that both past and future contexts help select the correct tag.

[6] For RNN models, only one (same) system is used to tag German, Greek and Spanish.

Table 1. *Token-level POS tagging accuracy for two baseline methods –B/L– (simple projection and URNN using MultiVec bilingual word embeddings as input), RNN*[6]*–Our–, Projection+RNN –Comb.– and state-of-the-art methods –SOTA– (Das & Petrov (2011), Duong et al. (2013) and Gouws & Søgaard (2015))*

Lang. / Model		French		German		Greek		Spanish	
		All	OOV	All	OOV	All	OOV	All	OOV
B/L	Simple projection	80.3	77.1	78.9	73.0	77.5	72.8	80.0	79.7
	URNN MultiVec	75.0	65.4	70.3	68.8	71.1	65.4	73.4	62.4
Our	URNN	78.5	70.0	76.1	76.4	75.7	70.7	78.8	72.6
	BRNN	80.6	70.9	77.5	76.6	77.2	71.0	80.5	73.1
	BRNN – OOV	81.4	77.8	77.6	77.8	77.9	75.3	80.6	74.7
Comb.	Proj. + URNN	84.5	78.8	81.5	77.0	78.3	74.6	83.6	81.2
	Proj. + BRNN	85.2	79.0	81.9	77.1	79.2	75.0	84.4	81.7
	Proj. + BRNN – OOV	**85.6**	**80.4**	82.1	**78.7**	79.9	**78.5**	**84.4**	**81.9**
SOTA	(Das 2011)	...	...	82.8	...	**82.5**	...	84.2	...
	(Duong 2013)	...	...	**85.4**	...	80.4	...	83.3	...
	(Gouws 2015)	...	...	84.8	...	...	...	82.6	...

The OOV rate of the test corpora is around twenty per cent.

Table 1 also shows the performance before and after performing our procedure for handling OOVs in BRNNs. It is shown that after replacing OOVs by the closest words using CBOW, the tagging accuracy significantly increases (McNemar's Test $p < 0.005$ on French, German and Greek).

As shown in the same table, our RNN models accuracy is close to that of the simple projection tagger. It achieves comparable results to Das and Petrov (2011) and Duong *et al.* (2013) (who used the full Europarl corpus while we use only a 65,000 subset of it) and to Gouws and Søgaard (2015) (who used extra resources such as Wiktionary and Wikipedia). Interestingly, RNN models learned using our common words representation (Section 3.1) seem to perform significantly better than RNN models using MultiVec bilingual word embeddings.

It is also important to note that only one single URNN and BRNN tagger applies to German, Greek and Spanish, so this is an authentic multilingual POS tagger. Finally, as for several other NLP tasks such as language modeling (where standard and NN-based models can be combined in order to obtain optimal results), the combination of standard and RNN-based approaches (*Projection+_*) seems necessary to further optimize POS tagging accuracies.

In order to know in what respect considering right context (bidirectional architecture) improves RNN model accuracy, we analyzed the French test corpus. In the example provided in Table 2, future time steps (tags of the words — de la Commission Européenne –) information helps to resolve the French word *Finances* tag ambiguity. We hypothesize that the context information is better represented in BRNN.

Table 2. *Effect of bidirectional architecture*

English	*Financial* situation of the European Parliament.
French	*Finances* de la Commission Européenne.
URNN	*Finances* / **VERB** ...
BRNN	*Finances* / **NOUN** ...

Table 3. *Word order divergence — unambiguous tag word –.*

EN Supervised Treetagger	... other/ADJ specific/ADJ groups/NOUN ...
FR Unsupervised URNN	... autres/ADJ groupes/NOUN spécifiques/ADJ ...

In case of word order divergence, we observed that our model can still handle some divergence, notably for the following cases:

- Obviously, if the current tag word is unambiguous (case of ADJ and NOUN order from English to French — see Table 3), then the context (RNN history) information has no effect.
- When the context is erroneous (due to the fact that word order for the target test corpus is different from the source training corpus), the right word tag can be recovered using the combination (RNN+Cross-lingual projection — see Table 4).

To deal with the word order divergence limitation, we also propose light supervision (adaptation) of RNN model. In Table 5, we report the results obtained after adaptation with a gradually increasing amount of target language data annotated (from 100 to 10,000 utterances).

We focus on German target language only. It is compared with two supervised approaches based on TnT or RNN. The supervised approaches are trained on the adaptation data only. For supervised RNN, it is important to mention that the input vector representation has a different dimension for each amount of adaptation data (we recall that the vector representation is $V_{wi}, i = 1, \ldots, N$, where N is the number of sentences, and N is growing from 100 to 10,000). The results show that our adaptation, on top of the unsupervised RNN is efficient in very low resource settings (<1,000 target language utterances). When more data is available (>1,000 utterances), the supervised approaches start to improve (but RNN and TnT are still complementary since their combination improves the tag accuracy).

Figure 4 details the behavior of the same methods for OOV words. We clearly see the limitation of the unsupervised URNN + adaptation to handle OOV words,

Table 4. *Word order divergence — ambiguous tag word -*

EN supervised treetagger	... two/NUM local/ADJ groups/NOUN ...
FR unsupervised URNN	... deux/NUM groupes/NOUN locaux/**NOUN** ...
Projection + URNN	... deux/NUM groupes/NOUN locaux/ **ADJ** ...

Table 5. *Lightly supervised model: effect of German adaptation corpus (manually annotated) size on unsupervised RNN model (unsupervised RNN + DE adaptation trained on EN Europarl and adapted to German)*

Model \ DE corpus size	0	100	500	1k	2k	5k	7k	10k
Unsupervised URNN + DE adaptation	76.1	**82.1**	**87.3**	**90.4**	90.7	91.2	91.4	92.4
Supervised URNN DE only	...	71.0	76.4	82.1	90.6	93.0	94.2	95.2
Supervised TnT DE only	...	80.5	86.5	89.0	92.2	94.1	95.3	95.7
Supervised URNN + Supervised TnT DE	...	81.0	86.7	90.1	**94.2**	**95.3**	**95.7**	**96.0**

Contrastive experiments with German supervised POS taggers using same data (RNN, TnT and RNN+TnT). 0 means no German corpus used during training.

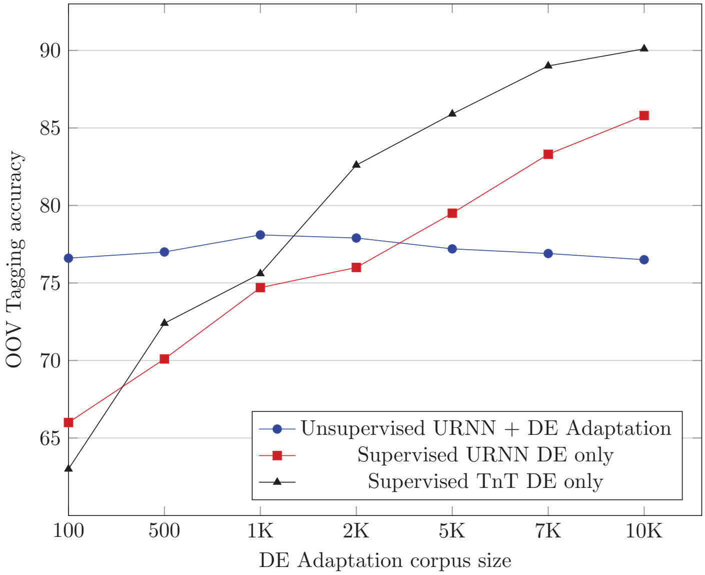

Fig. 4. (Colour online) Accuracy on OOV according to German training corpus size for unsupervised URNN + DE adaptation, supervised URNN DE and supervised TNT DE.

since the input vector representation is the same (comes from the initial parallel corpus) and does not evolve as more German adaptation data is available.

In order to know the impact of adaptation, we analyzed the German test corpus. The example provided in Table 6, shows that RNN better handles tag ambiguity with adaptation: In the unsupervised RNN model, the word *kandidiert* is tagged as a noun (NOUN), whereas it is a verb (VERB) in this particular context.

5 Multilingual super sense tagging-based RNN

Our RNN model is applied to a more complex task: multilingual SST. In order to measure the impact of the parallel corpus quality on our method, we also learn our

Table 6. *Effect of adaptation on German data*

German	... er denn kandidiert
Unsupervised URNN	... er/PRON denn/ADV kandidiert/**NOUN**
Unsupervised URNN +DE adapt.(1k)	... er/PRON denn/ADV kandidiert/**VERB**

SST models using the multilingual parallel corpus MultiSemCor (MSC) which is the result of manual or automatic translation of SemCor from English into Italian and French.

5.1 Super sense tagset

SST is an NLP task that consists of annotating each significant entity in a text, within a general semantic taxonomy defined by the WordNet (Fellbaum 1998) lexicographer classes (called super-senses). In Ciaramita and Altun (2006), SST was defined as a task half-way between named entity recognition and Word Sense Disambiguation (WSD): It is an extension of NER, since it uses a larger set of semantic categories, and it is an easier and more practical task with respect to WSD, that deals with very specific senses (the complete list of super-senses and a short description is shown in Table 7).

5.2 RNN variants

We propose three new RNN variants to take into account low-level (POS) information in a higher level (SST) annotation task. The question addressed here is at which layer of the RNN should this low-level information be included to improve SST performance? As specified in Figure 5, the POS information can be introduced either at input layer or at forward layer (forward and backward layers for BRNN) or at compression layer. In all these RNN variants, the POS of the current word is also represented with a vector (POS(t)). Its dimension corresponds to the number of POS tags in the tagset (universal tagset of Petrov *et al.* (2012) is used). We propose one *hot* vector representation where only one value is set to 1 and corresponds to the index of current tag (all other values are 0).

5.3 Data

SemCor. The SemCor (Miller *et al.* 1993)[7] is a subset of the Brown Corpus (Kucera and Francis 1979) labeled with the *WordNet* (Fellbaum 1998) senses. In Table 8, we provide an example extracted from SemCor.

MultiSemCor. The English–Italian MultiSemcor (MSC-IT-1) corpus is a manual translation of the English SemCor to Italian (Bentivogli, Forner and Pianta 2004)[8]. As we already mentioned, we are also interested in measuring the impact of the

[7] `http://web.eecs.umich.edu/~mihalcea/downloads.html#semcor`
[8] `http://multisemcor.fbk.eu/`

Table 7. *Nouns and verbs super sense labels (WordNet lexicographer classes), and short description (from the Wordnet documentation)*

NOUNS			
SuperSense	NOUNS DENOTING	SuperSens	NOUNS DENOTING
Act	Acts or actions	Object	Natural objects (not man-made)
Animal	Animals	Quantity	Quantities and units of measure
Artifact	Man-made objects	Phenomenon	Natural phenomena
Attribute	Attributes of people and objects	Plant	Plants
Body	Body parts	Possession	Possession and transfer of possession
Cognition	Cognitive processes and contents	Process	Natural processes
Communication	Communicative processes and contents	Person	People
Event	Natural events	Relation	Relations between people or things or ideas
Feeling	Feelings and emotions	Shape	Two- and three-dimensional shapes
Food	Foods and drinks	State	Stable states of affairs
Group	Groupings of people or objects	Substance	Substances
Location	Spatial position	Time	Time and temporal relations
Motive	Goals	Tops	Abstract terms for unique beginners

VERBS			
SuperSense	VERBS OF	SuperSens	VERBS OF
Body	Grooming, dressing and bodily care	Emotion	Feeling
Change	Size, temperature change, intensifying	Motion	Walking, flying, swimming
Cognition	Thinking, judging, analyzing, doubting	Perception	Seeing, hearing, feeling
Communication	Telling, asking, ordering, singing	Possession	Buying, selling, owning
Competition	Fighting, athletic activities	Social	Political and social activities and events
Consumption	Eating and drinking	Stative	Being, having, spatial relations
Contact	Touching, hitting, tying, digging	Weather	Raining, snowing, thawing, thundering
Creation	Sewing, baking, painting, performing		

Table 8. *An example of SemCor (v.3.0) in SGML format (standard generalized markup language)*

```
<contextfile concordance=brown>
  <context filename=br-a01 paras=yes>
  [...]
   <s snum=29>
    <wf cmd=ignore pos=DT>The</wf>
    <wf cmd=done pos=NN lemma=couple wnsn=2 lexsn=1:14:01:: >couple</wf>
    <wf cmd=done pos=VBD ot=notag>was</wf>
    <wf cmd=done pos=VB lemma=marry wnsn=1 lexsn=2:41:00::>married</wf>
    <wf cmd=done rdf=aug pos=NN lemma=aug wnsn=1 lexsn=1:28:00::>Aug.</wf>
    <wf cmd=done pos=CD ot=notag>2</wf>
    <punc>,</punc>
    <wf cmd=done pos=CD ot=notag>1913</wf>
    <punc>.</punc>
   </s>
  [...]
  </context>
</contextfile>
```

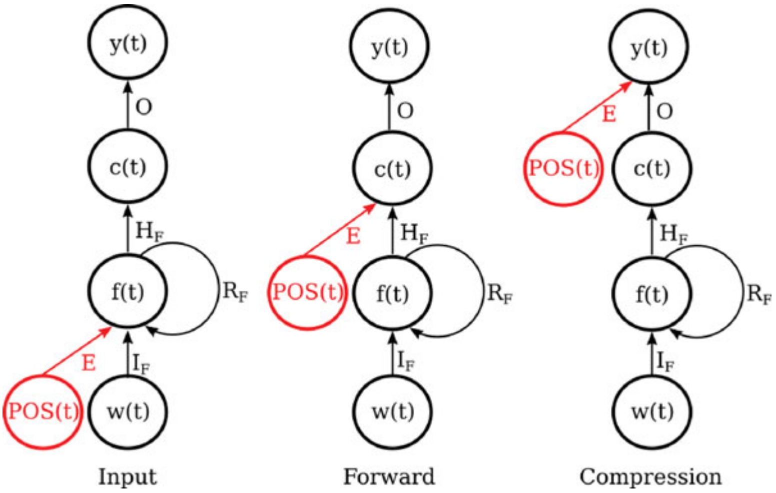

Fig. 5. (Colour online) URNN variants with POS information at three levels: (a) input layer, (b) forward layer, (c) compression layer.

parallel corpus quality on our method. For this, we use two translation systems: (a) Google Translate to translate the English SemCor to Italian (MSC-IT-2) and French (MSC-FR-2); (b) LIG machine translation phrase based system, off-the-shelf SMT system based on Moses (Koehn *et al.* 2007) taken from LIG (Besacier *et al.* 2012) to translate the English SemCor to French (MSC-FR-1)[9] (Salah *et al.* 2016).

[9] https://github.com/getalp/WSD-TALN2016-Hadjsalahetal

Table 9. *WordNet synset, BabelNet sense and super sense mapping*

Word	WordNet synset	BabelNet sense	Super sense
Couple	couple%1:14:01::	bn:00023269n	noun.group
Married	marry%2:41:00::	bn:00085614v	verb.social
Aug.	aug%1:28:00::	bn:00007140n	noun.time

Training corpus. The SemCor was labeled with the *WordNet* synsets. However, because we train models for SST, we convert SemCor synsets annotations to super senses. We train our models using the four different versions of MSC (MSC-IT-1,2 - MSC-FR-1,2), with modified Semcor on source side.

Test Corpus. To evaluate our models, we used the SemEval 2013 Task 12 (multilingual WSD) (Navigli, Jurgens and Vannella 2013) test corpora, which are available in five languages (English, French, German, Spanish and Italian) and labeled with *BabelNet* (Navigli and Ponzetto 2012) senses. We map BabelNet senses to WordNet synsets, then WordNet synsets are mapped to super senses (see Tabel 9). We used the sense mapping between Princeton WordNet 3.0 and BabelNet 1.1.1 provided by Nasiruddin *et al.* (2015).

5.3.1 SST systems evaluated

The goals of our SST experiments are the following: First, to investigate the effectiveness of using POS information to build multilingual super sense tagger, second to measure the impact of the parallel corpus quality (manual or automatic translation) on our RNN models (URNN, BRNN and our proposed variants). To summarize, we build four super sense taggers based on direct cross-lingual projection (see Section 2.2.1) using four versions of MSC (MSC-IT-1, MSC-IT-2, MSC-FR-1 and MSC-FR-2) described above. Then we use the same four versions to train our multilingual SST models based on URNN and BRNN. For learning our multilingual SST models based on RNN variants proposed in Section 5.1, we also tag SemCor using *TreeTagger* (POS tagger proposed by Schmid (1995)).

5.4 Results and discussion

5.4.1 Results

Our models are evaluated on SemEval 2013 Task 12 test corpora. Results are directly comparable with those of systems which participated to this evaluation campaign. We report two SemEval 2013 (unsupervised) system results for comparison:

- **MFS Semeval 2013**: The most frequent sense is the baseline provided by SemEval 2013 for Task 12, this system is a strong baseline, which is obtained by using an external resource (the WordNet most frequent sense).

Table 10. *Super sense tagging (SST) accuracy for simple Projection, RNN and their combination*

	Lang. Model	Italian		French	
		MSC-IT-1 trans man.	MSC-IT-2 trans. auto	MSC-FR-1 trans. auto	MSC-FR-2 trans auto.
B/L	Simple Projection	61.3	45.6	42.6	44.5
Our SST Based RNN	URNN	59.4	46.2	46.2	47.0
	BRNN	59.7	46.2	46.0	47.2
	URNN-POS-In	61.0	47.0	46.5	47.3
	URNN-POS-H1	59.8	46.5	46.8	47.4
	URNN-POS-H2	63.1	48.7	47.7	49.8
	BRNN-POS-In	61.2	47.0	46.4	47.3
	BRNN-POS-H1	60.1	46.5	46.8	47.5
	BRNN-POS-H2	63.2	48.8	47.7	50
	BRNN-POS-H2 - OOV	64.6	49.5	48.4	50.7
Combination	Proj. + URNN	62.0	46.7	46.5	47.4
	Proj. + BRNN	62.2	46.8	46.4	47.5
	Proj. + URNN-POS-In	62.9	47.4	46.9	47.7
	Proj. + URNN-POS-H1	62.5	47.0	47.1	48.0
	Proj. + URNN-POS-H2	63.5	49.2	48.0	50.1
	Proj. + BRNN-POS-In	62.9	47.5	46.9	47.8
	Proj. + BRNN-POS-H1	62.7	47.0	47.0	48.0
	Proj. + BRNN-POS-H2	63.6	49.3	48.0	50.3
	Proj. + BRNN-POS-H2 - OOV	**64.7**	49.8	48.6	51.0
S-E	MFS Semeval 2013	60.7		**52.4**	
	GETALP (Schwab *et al.* 2012)	40.2		34.6	

- **GETALP**: A fully unsupervised WSD system proposed by Schwab *et al.* (2012) based on Ant-Colony algorithm.

The DAEBAK! (Manion and Sainudiin 2013) and the UMCC-DLSI systems (Gutiérrez Vázquez *et al.* 2011) have also participated to SemEval 2013 Task 12. However, they use a supervised approach.[10]

Table 10 shows the results obtained by our RNN models and by two SemEval 2013 WSD systems. URNN-POS-X and BRNN-POS-X refer to our RNN variants: *In* means input layer, *H1* means first hidden layer and *H2* means second hidden layer. We achieve the best performance on Italian using MSC-IT-1 clean corpus, while noisy training corpus degrades SST performance. The best results are obtained with

[10] DAEBAK! and UMCC-DLSI for SST have obtained: 68.1 per cent and 72.5 per cent on Italian; 59.8 per cent and 67.6 per cent on French.

Table 11. *Improved tagged example for French target language*

English	... who also serves as the regional Mexico climate_change *adviser*.
French	... qui est également *conseiller* sur le climat pour le Mexique.
BRNN	... *conseiller* / **verb.communication** ...
BRNN-POS-H2	... *conseiller* / **noun.person** ...

combination of simple projection and RNN which confirms (as for POS tagging) that both approaches are complementary.

We also observe that the RNN approach seems more robust than simple projection on noisy corpora. This is probably due to the fact that no word alignments are required in our cross language RNN. Finally, BRNN-POS-H2-OOV achieves the best performance, which shows that the integration of POS information in RNN models and dealing with OOV words are useful to build efficient multilingual super senses taggers. Finally, it is worth mentioning that integrating low-level (POS) information lately (last hidden layer) seems to be the best option in our case.

The example in Table 11, shows that the integration of POS information in our neural SST tagger (BRNN-POS-H2) helps to resolve the French word *conseiller* (adviser in English) SST tag ambiguity, tagged by BRNN (model without the integration of POS information) as *verb.communication*, whereas it is an *noun.person*

6 Conclusion

In this article, we have described an approach based on RNNs to induce multilingual NLP tools. In particular, we have used URNN and BRNN architectures on two NLP tasks: POS tagging and SST. We have also proposed and experimented new RNN variants which take into account low-level information (POS) in a higher level task (SST). Our approach needs only parallel corpora without using word alignment information and it does not assume any knowledge about target languages (however, one requirement is that the two languages (source and target) are not too syntactically divergent).

We first empirically evaluated the proposed approach on two unsupervised POS taggers: (1) English–French bilingual POS tagger; and (2) English–German–Greek–Spanish multilingual POS tagger. The performance of the second model is close to state-of-the-art with only a subset (65, 000 sentences) of Europarl corpus used (when state-of-the-art approaches use the whole Europarl corpus). In addition, when a small amount of supervised data is available, the experimental results demonstrated the effectiveness of our approach in a weakly supervised context (well adapted to low resource scenarios).

Second, in order to demonstrate the genericity of our approach we have applied our RNN model on a more complex task: multilingual SST. We have investigated two pairs of languages: English–Italian and English–French which allowed us to measure the impact of the parallel corpus quality on the results of our approach. For the English–Italian pair, we have used two MSC parallel corpora: The first results from English–Italian manual translation and the second was constructed using English–Italian automatic translation. Concerning the English–French pair, we have used two MSC parallel corpora generated with two Statistical Machine Translation tools. The obtained results (on SST) showed that our approach seems more robust than state-of-the-art methods (based on word alignment) on noisy corpora. Additionally, we have observed that the integration of POS information in RNN models is useful to build efficient multilingual super senses taggers.

Finally, we have also proposed a method to deal with OOV words which has led to improvements for both tasks (multilingual POS tagging and multilingual SST).

We then conclude that our approach is generic and has the following advantages: (a) it uses a language-independent word representation (based only on word co-occurrences in a parallel corpus), (b) it provides authentic multilingual taggers (one tagger for *N* languages) and (c) it can be easily adapted to a new target language (when a small amount of supervised data is available).

For future work, we plan to apply multitask learning to build systems that simultaneously perform syntactic and semantic analysis. We also plan, on the one hand, to exploit multiple source languages to improve our RNN taggers with our common (multilingual) vector representation (this is possible with a multi-parallel corpus and our common words representation), and on the other hand, to apply our approach to truly under-resourced languages as defined in Besacier *et al.* (2014). In the case of strong differences in word order between source and target languages, and based on Aufrant, Wisniewski and Yvon (2016) works, we plan to use pre-processing based on features extracted from large typological databases such as World Atlas of Language Structures[11] for instance.

References

Al-Rfou, R., Perozzi, B., and Skiena, S. 2013. Polyglot: distributed word representations for multilingual nlp. In *Proceedings of the SIGNLL Conference on Computational Natural Language Learning*, pp. 183–192.

Annesi, P., and Basili, R. 2010. Cross-lingual alignment of FrameNet annotations through Hidden Markov Models. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, Berlin, Heidelberg, pp. 12–25.

Aufrant, L., Wisniewski, G., and Yvon, F. 2016. Zero-resource dependency parsing: boosting delexicalized cross-lingual transfer with linguistic knowledge. In *Proceedings of the 26th International Conference on Computational Linguistics*, pp. 119–130.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. 2003. A neural probabilistic language model. *Journal of Machine Learning Research* **3**, 1137–1155.

Bengio, Y., Schwenk, H., Senécal, J.-S., Morin, F., and Gauvain, J.-L. 2006. Neural probabilistic language models. In H. Dawn E. and J. Lakhmi C. (eds.), *Innovations in Machine Learning*, pp. 137–186. Berlin, Heidelberg: Springer.

Bentivogli, L., Forner, P., and Pianta, E. 2004. Evaluating cross-language annotation transfer in the multisemcor corpus. In *Proceedings of the 20th International Conference on Computational Linguistics*, Association for Computational Linguistics, pp. 364–371.

[11] http://wals.info

Bérard, A., Servan, C., Pietquin, O, and Besacier, L. 2016. MultiVec: a multilingual and multilevel representation learning toolkit for NLP. In *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference*, pp. 4188–4192.

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. 2014. Automatic speech recognition for under-resourced languages: a survey. *Speech Communication* **56**: 85–100.

Besacier, L., Lecouteux, B., Azouzi, M., and Luong, N.-Q. 2012. The LIG English to French machine translation system for IWSLT 2012. In *Proceedings of the 9th International Workshop on Spoken Language Translation*, pp. 102–108.

Brants, T. 2000. TnT: a statistical part-of-speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, Association for Computational Linguistics, pp. 224–231.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics* **19**: 263–311.

Buchholz, S., and Marsi, E. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning*, Association for Computational Linguistics, pp. 149–164.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. 2014. On the properties of neural machine translation: encoder–decoder approaches. In *Proceedings of the Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111.

Ciaramita, M., and Altun, Y. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 594–602.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* **12**: 2493–2537.

Das, D., and Petrov, S. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, Association for Computational Linguistics, pp. 600–609.

Duong, L., Cook, P., Bird, S., and Pecina, P. 2013. Simpler unsupervised POS tagging with bilingual projections. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 2, pp. 634–639.

Durrett, G., Pauls, A., and Klein, D. 2012. Syntactic transfer using a bilingual lexicon. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, pp. 1–11.

Elman, J. L. 1990. Finding structure in time. *Cognitive science* **14**: 179–211.

Fellbaum, C. 1998. *WordNet*. Wiley Online Library, Cambridge, MA: MIT Press.

Fraser, A., and Marcu, D. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics* **33**: 293–303.

Garside, R., Leech, G. N., and McEnery, T. 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Taylor & Francis, Abingdon.

Gouws, S., and Søgaard, A. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 14th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1386–1390.

Gouws, S., Bengio, Y., and Corrado, G. 2015. BilBOWA: fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 748–756.

Graves, A. 2012. Supervised sequence labelling. In *Supervised Sequence Labelling with Recurrent Neural Networks*, pp. 5–13. Berlin, Heidelberg: Springer.

Gutiérrez Vázquez, Y., Fernández Orquín, A., Montoyo Guijarro, A., Vázquez Pérez, S. 2011. *Enriching the Integration of Semantic Resources Based on Wordnet*. Sociedad Española para el Procesamiento del Lenguaje Natural, **47**: 249–257, Huelva, Spain.

Henderson, J. 2004. Discriminative training of a neural network statistical parser. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, pp. 95–102.

Jiang, W., Liu, Q., and Lü, Y. 2011. Relaxed cross-lingual projection of constituent syntax. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 1192–1201.

Jiang, W., Lü, Y., Huang, L., and Liu, Q. 2015. Automatic adaptation of annotations. *Computational Linguistics Journal* **41**: 119–147.

Kim, S., Toutanova, K., and Yu, H. 2012. Multilingual named entity recognition using parallel data and metadata from wikipedia. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics,* vol. 1, pp. 694–702.

Koehn, P. 2005. Europarl: a parallel corpus for statistical machine translation. *MT Summit* **5**: 79–86.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., and Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Association for Computational Linguistics, pp. 177–180.

Kucera, H., and Francis, W. 1979. *A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers* (Revised and amplified from 1967 version). Providence, RI: Brown University Press.

Li, S., Graça, J. V., and Taskar, B. 2012. Wiki-ly supervised part-of-speech tagging. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, pp. 1389–1398.

Luong, T., Pham, H., and Manning, C. D. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pp. 151–159.

Manion, S. L., and Sainudiin, R. 2013. DAEBAK!: peripheral diversity for multilingual word sense disambiguation. In *Proceedings of SemEval*, pp. 250–254.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., and Khudanpur, S. 2010. Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, pp. 1045–1048.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Advances in Neural Information Processing Systems*, pp. 3111–3119.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. 1993. A semantic concordance. In *Proceedings of the Workshop on Human Language Technology*, Association for Computational Linguistics, pp. 303–308.

Nasiruddin, M., Tchechmedjiev, A., Blanchon, H., and Schwab, D. 2015. Création rapide et efficace dun système de désambiguïsation lexicale pour une langue peu dotée. In *Proceedings of the 22nd TALN (Traitement Automatique des Langues Naturelles) Conference.*

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., Jurgens, D., and Vannella, D. 2013. Semeval-2013: Multilingual word sense disambiguation. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, vol. 2, pp. 222–231.

Och, F. J., and Ney, H. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, pp. 440–447.

Pado, S., and Pitel, G.. 2007. Annotation précise du français en sémantique de rôles par projection cross-linguistique. In *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (communications orales)*, pp. 271–280.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* **22**: 1345–1359.

Passban, P., Liu, Q., and Way, A. 2017. Translating low-resource languages by vocabulary adaptation from close counterparts. *ACM Transactions on Asian and Low-Resource Language Information Processing* **16**: 29.

Petrov, S., Das, D., and McDonald, R. 2012. A universal part-of-speech tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, European Language Resources Association, pp. 2089–2096.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1985. Learning internal representations by error propagation. DTIC Document. No. ICS-8506. California Univ San Diego La Jolla Inst for Cognitive Science.

Salah, M. H., Blanchon, H., Zrigui, M., and Schwab, D. 2016. Amélioration de la traduction automatique dun corpus annoté. In *Proceedings of the 23rd TALN (Traitement Automatique des Langues Naturelles) Conference*.

Schmid, H. 1995. Treetagger — a language independent part-of-speech tagger. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, vol. 46, p. 28. Available at `https://protect-eu.mimecast.com/s/STrqCK8y8fB91wiMedpW?domain=cis.uni-muenchen.de` `http://www.cis.uni-muenchen.de/∼schmid/tools/TreeTagger/`

Schmidhuber, J. 1992. A fixed size storage O (n3) time complexity learning algorithm for fully recurrent continually running networks. *Neural Computation* **4**: 243–248.

Schuster, M., and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* **45**: 2673–2681.

Schwab, D., Goulian, J., Tchechmedjiev, A., and Blanchon, H. 2012. Ant colony algorithm for the unsupervised word sense disambiguation of texts: comparison and evaluation. In *Proceedings of the 25th International Conference on Computational Linguistics*, pp. 2389–2404.

Sundermeyer, M., Oparin, I., Gauvain, J.-L., Freiberg, B., Schluter, R., and Ney, H. 2013. Comparison of feedforward and recurrent neural network language models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8430–8434.

Sutskever, I., Vinyals, O., and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 3104–3112.

Täckström, O., McDonald, R., and Uszkoreit, J. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, pp. 477–487.

Täckström, O., McDonald, R., and Nivre, J. 2013. Target language adaptation of discriminative transfer parsers. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, pp. 1061–1071.

Täckström, O., Das, D., Petrov, S., McDonald, R., and Nivre, J. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics* **1**: 1–12.

Titov, I., and Klementiev, A. 2012. Crosslingual induction of semantic roles. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 647–656.

Van der Plas, L., and Apidianaki, M. 2014. Cross-lingual word sense disambiguation for predicate labelling of french. In *Proceedings of the 21st TALN (Traitement Automatique des Langues Naturelles) Conference*, pp. 46–55.

Veronis, J. 2000. Annotation automatique de corpus: panorama et état de la technique. *Ingénierie des langues* **4**(4): 111–129.

Veronis, J., Hamon, O., Ayache, C., Belmouhoub, R., Kraif, O., Laurent, D., Nguyen, T. M. H., Semmar, N., Stuck, F., and Zaghouani, W. 2008. Arcade II Action de recherche concertée sur l'alignement de documents et son évaluation. Chapitre2, Editions Hermés.

Van der Maaten, L., and Hinton, G. (2008) Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**: 2579–2605.

Wisniewski, G., Pécheux, N., Gahbiche-Braham, S., and Yvon, F. 2014. Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* , vol. 14, pp. 1779–1785.

Yarowsky, D., Ngai, G., and Wicentowski, R. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the 1st International Conference on Human Language Technology Research*, pp. 1–8.