Unsupervised and Lightly Supervised Part-of-Speech Tagging Using Recurrent Neural Networks

Othman Zennaki [1,2] **Nasredine Semmar** [1] **Laurent Besacier** [2]

[1]CEA, LIST, Vision and Content Engineering Laboratory, Gif-sur-Yvette, France
{othman.zennaki, nasredine.semmar}@cea.fr
[2]Laboratory of Informatics of Grenoble, Univ. Grenoble-Alpes, Grenoble, France
laurent.besacier@imag.fr

Abstract

In this paper, we propose a novel approach to induce automatically a Part-Of-Speech (POS) tagger for resource-poor languages (languages that have no labeled training data). This approach is based on cross-language projection of linguistic annotations from parallel corpora without the use of word alignment information. Our approach does not assume any knowledge about foreign languages, making it applicable to a wide range of resource-poor languages. We use Recurrent Neural Networks (RNNs) as multilingual analysis tool. Our approach combined with a basic cross-lingual projection method (using word alignment information) achieves comparable results to the state-of-the-art. 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

Nowadays, Natural Language Processing (NLP) tools (part-of-speech tagger, sense tagger, syntactic parser, named entity recognizer, semantic role labeler, etc.) with the best performance are those built using supervised learning approaches for resource-rich languages (where manually annotated corpora are available) such as English, French, German, Chinese and Arabic. However, for a large number of resource-poor languages, annotated corpora do not exist. Their manual construction is labor intensive and very expensive, making supervised approaches not feasible.

The availability of parallel corpora has recently led to several strands of research work exploring the use of unsupervised approaches based on linguistic annotations projection from the (resource-rich) *source* language to the (under-resourced) *target* language. 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 NLP tools for these languages. Unfortunately, the state-of-the-art in unsupervised methods, is still quite far from supervised learning approaches. For example, Petrov et al. (2012) obtained an average accuracy of 95.2% for 22 resource-rich languages supervised POS taggers, while the state-of-the-art in the unsupervised POS taggers achieved by Das and Petrov (2011) and Duong et al. (2013) with an average accuracy reaches only 83.4% on 8 European languages. Section 2 presents a brief overview of related work.

In this paper, we first adapt a similar method than the one of Duong et al. (2013) [1], to build an unsupervised POS tagger based on a simple cross-lingual projection (Section 3.1). Next, we explore the possibility of using a recurrent neural network (RNN) to induce multilingual NLP tools, without using word alignment information. To show the potential of our approach, we firstly investigate POS tagging.

In our approach, a parallel corpus between a resource-rich language (having a POS tagger) and a lower-resourced language is used to extract a common words representation (cross-lingual words representation) based only on sentence level alignment. This representation is used with the source side of the parallel corpus (tagged corpus) to learn a neural network POS tagger for the source language. No

[1]We did not use incremental training (as Duong et al. (2013) did).

29th Pacific Asia Conference on Language, Information and Computation pages 133 - 142
Shanghai, China, October 30 - November 1, 2015

word alignment information is needed in our approach. Based on this common representation of source and target words, this neural network POS tagger can also be used to tag target language text (Section 3.2).

We assume that these two models (baseline cross-lingual projection and RNN) are complementary to each other (one relies on word-alignment information while the other does not), and the performance can be further improved by combining them (linear combination presented in Section 3.3). This unsupervised RNN model, obtained without any target language annotated data, can be easily adapted in a weakly supervised manner (if a small amount of annotated target data is available) in order to take into account the target language specificity (Section 4).

To evaluate our approach, we conducted an experiment, which consists of two parts. First, using only parallel corpora, we evaluate our unsupervised approach for 4 languages: French, German, Greek and Spanish. Secondly, the performance of our approach is evaluated for German in a weakly supervised context, using several amounts of target adaptation data (Section 5). Finally, Section 6 concludes our study and presents our future work.

2 Related Work

Several studies have used cross-lingual projection to transfer linguistic annotations from a resource-rich language to a resource-poor language in order to train NLP tools for the target language. The projection approach has been successfully used to transfer several linguistic annotations between languages. Examples include POS (Yarowsky et al., 2001; Das and Petrov, 2011; Duong et al., 2013), named entity (Kim and Lee, 2012), syntactic constituent (Jiang et al., 2011), word senses (Bentivogli et al., 2004; Van der Plas and Apidianaki, 2014), and semantic role labeling (Padó , 2007; Annesi and Basili, 2010).

In these approaches, the source language is tagged, and tags are projected from the source language to the target language through the use of word alignments in parallel corpora. Then, these partial noisy annotations can be used in conjunction with robust learning algorithms to build unsupervised NLP tools. One limitation of these approaches is due to the poor accuracy of word-alignment algorithms, and also to the weak or incomplete inherent match between the two sides of a bilingual corpus (the alignment is not only a one-to-one mapping, it can also be one-to-many, many-to-one, many-to-many or some words can remain unaligned). 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 et al. (2012), Täckström et al. (2013b) and Wisniewski et al. (2014) have proposed to improve projection performance by using a dictionary of valid tags for each word (coming from Wiktionary [2]).

In another vein, various studies based on cross-lingual representation learning methods have proposed to avoid using such pre-processed and noisy alignments for label projection. First, these approaches learn language-independent features, across many different languages (Al-Rfou et al., 2013). Then, the induced representation space is used to train NLP tools by exploiting labeled data from the source language and apply them in the target language. To induce interlingual features, several resources have been used, including bilingual lexicon (Durrett et al., 2012; Gouws and Søgaard, 2015a) and parallel corpora (Täckström et al., 2013a; Gouws et al., 2015b). Cross-lingual representation learning have achieved good results in different NLP applications such as cross-language POS tagging and cross-language super sense (SuS) tagging (Gouws and Søgaard, 2015a), cross-language named entity recognition (Täckström et al., 2012), cross-lingual document classification and lexical translation task (Gouws et al., 2015b), cross language dependency parsing (Durrett et al., 2012; Täckström et al., 2013a; Xiao and Guo, 2014) and cross language semantic role labeling (Titov and Klementiev, 2012). Our approach described in next section, is inspired by these works since we also try to learn a common language-independent feature space. Our common (multilingual) representation is based on the occurrence of source and target words in a parallel corpus. Using this representation, 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

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

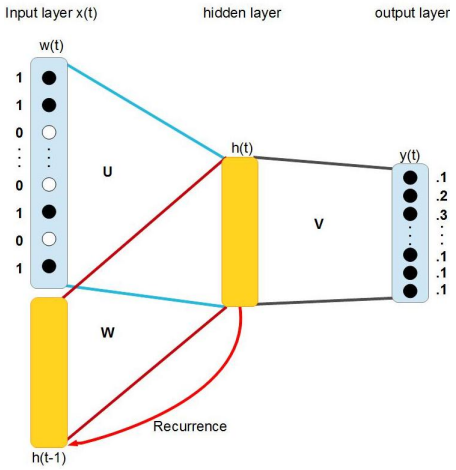

Figure 1: Architecture of the recurrent neural network.

labeled text and apply it to tag target language text. We also show that the architecture proposed is well suited for lightly supervised training (adaptation).

Finally, several works have investigated how to apply neural networks to NLP applications (Bengio et al., 2006; Collobert and Weston, 2008; Collobert et al., 2011; Henderson, 2004; Mikolov et al., 2010; Federici and Pirrelli, 1993). While Federici and Pirrelli (1993) was one of the earliest attempts to develop a part-of-speech tagger based on a special type of neural network, Bengio et al. (2006) and Mikolov et al. (2010) applied neural networks to build language models. Collobert and Weston (2008) and Collobert et al. (2011) employed a deep learning framework for multi-task learning including part-of-speech tagging, chunking, named-entity recognition, language modelling and semantic role-labeling. Henderson (2004) proposed training methods for learning a statistical parser based on neural network.

3 Unsupervised Approach Overview

To avoid projecting label information from deterministic and error-prone word alignments, we propose to represent the bilingual word alignment information intrinsically in a neural network architecture. The idea consists in implementing a neural network as a cross-lingual POS tagger and show that, in combination with a simple cross-lingual projection method, this achieves comparable results to state-of-the-art unsupervised POS taggers.

Our approach is the following: we assume that we have a POS tagger in the source language and a parallel corpus. The key idea is to learn a bilingual neural network POS tagger on the pre-annotated *source* side of the parallel corpus, and to use it for tagging *target* text. Before describing our bilingual neural network POS tagger, we present the simple cross-lingual projection method, considered as our baseline in this work.

3.1 Unsupervised POS Tagger Based on a Simple Cross-lingual Projection

Our simple POS tagger (described by Algorithm 1) is close to the approach introduced in Yarowsky et al. (2001). These authors were the first to use automatic word alignments (from a bilingual parallel corpus) to project annotations from a *source* language to a *target* language, to build unsupervised POS taggers. The algorithm is shortly recalled below.

Algorithm 1 : Simple POS Tagger

1: Tag source side of the parallel corpus.
2: Word align the parallel corpus with Giza++ (Och and Ney, 2000) or other word alignment tools.
3: Project tags directly for 1-to-1 alignments.
4: For many-to-one mappings project the tag of the middle word.
5: The unaligned words (target) are tagged with their most frequent associated tag in the corpus.
6: Learn POS tagger on target side of the bi-text with, for instance, TNT tagger (Brants, 2000).

3.2 Unsupervised POS Tagger Based on Recurrent Neural Network

There are two major architectures of neural networks: Feedforward (Bengio et al., 2006) and Recurrent Neural Networks (RNN) (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 recurrent neural networks do not use a context of limited size. This property led us to use, in our experiments, a simple recurrent architecture (Elman, 1990).

In this section, we describe in detail our method for building an unsupervised POS tagger for a target language based on a recurrent neural network.

3.2.1 Model description

The RNN consists of at least three layers: input layer in time t is $x(t)$, hidden layer $h(t)$ (also called context layer), and output layer is denoted as $y(t)$. All neurons of the input layer are connected to every neuron of hidden layer by weight matrix U and W. The weight matrix V connects all neurons of the hidden layer to every neuron of output layer, as it can be seen in Figure 1.

In our RNN POS tagger, the input layer is formed by concatenating vector representing current word w, and the copy of the hidden layer at previous time. We start by associating to each word in both the source and the target vocabularies a common vector representation, namely $V_{wi}, i = 1, ..., N$, where N is the number of parallel sentences (bi-sentences in the parallel corpus). If w appears in i-th bi-sentence of the parallel corpus then $V_{wi} = 1$. Therefore, all input neurons corresponding to current word w are set to 0 except those that correspond to bi-sentences containing w, which are set to 1. The idea is that, in general, a source word and its target translation appear together in the same bi-sentences and their vector representations are close.We can then use the RNN POS tagger, initially trained on source side, to tag the target side (because of our *common vector representation*).

We also use two hidden layers (our preliminary experiments have shown better performance than one hidden layer), with variable sizes (usually 80-1024 neurons) and sigmoid activation function. These hidden layers inherently capture word alignment information. The output layer of our model contains 12 neurons, this number is determined by the POS tagset size. To deal with the potential mismatch in the POS tagsets of source and target languages, we adopted the Petrov et al. (2012) universal tagset (12 tags common for 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 (all other categories, e.g., foreign words, abbreviations). Therefore, each output neuron corresponds to one POS tag in the tagset. The softmax activation function is used to normalize the values of output neurons to sum up to 1. Finally, the current word w (in input) is tagged with most probable output tag.

3.2.2 Training the model

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 POS tagger). Before training the model, the following pre-processing steps are performed :

- Source side of training corpus and validation corpus are annotated (using the available supervised POS tagger).
- Using a parallel corpus, we build the common vector representations for source and target side words.

Then, the neural network is trained through several epochs. Algorithm 2 below describes one training epoch.

Algorithm 2 : Training RNN POS Tagger

1: Initialize weights with Normal distribution.
2: Set time counter t = 0, and initialize state of the neurons in the hidden layer h(t) to 1.
3: Increase time counter t by 1.
4: Push at the input layer w(t) the vector representation of the current (source) word of training corpus.
5: Copy the state of the hidden layer h(t-1) to the input layer.
6: Perform a forward pass to obtain the predicted output y(t).
7: Compute the gradient of the error in the output layer $e_o(t) = d(t) - y(t)$ (difference between the predicted *y(t)* and the desired output *d(t)*).
8: Propagate the error back through the network and update weights with stochastic gradient descent using Back-Propagation *(BP)* and Back-Propagation-through-time *(BPTT)* (Rumelhartet al., 1985).
9: If not all training inputs were processed, go to 3.

After each epoch, the neural network is used to tag the validation corpus, then the result is compared with the result of the supervised POS tagger, to calculate the *per-token* accuracy. 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. After that, 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 $\alpha = 0.1$.

After learning the model, step 2 simply consists in using the trained model as a target language POS 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 European Commission* would be translated into *la Commission europenne*. In order to deal with the word order constraints, we combined the RNN model with the cross-lingual projection model, and we also propose Light Supervision (adaptation) of RNN model where a few amount of target data will help to learn the word order (and consequently POS order) in the target language.

3.3 Combining Simple Cross-lingual Projection and RNN Models

Since the simple cross-lingual projection model *M1* and RNN model *M2* use different strategies for POS tagging (TNT is based on Markov models while RNN is a neural network), we assume that these two models are complementary. In addition, model *M2* does not implement any out-of-vocabulary (OOV) words processing yet. So, 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

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

where, C_{M1} and C_{M2} are, respectively the context of w considered by *M1* and *M2*. The relative importance of each model is adjusted through the interpolation parameter μ.

The word w is tagged with the most probable tag, using the function f described as

$$f(w) = \arg\max_t (P_{M12}(t|w)) \quad (2)$$

4 Light Supervision (adaptation) of RNN model

While the unsupervised RNN 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) in order to capture target language specificity. Such an adaptation is performed on top of the unsupervised RNN model without retraining the full model. The full process is the following (steps 1 and 2 correspond to the unsupervised case):

1. Each word in the parallel corpus is represented by a binary occurrence vector (same initial common vector representation).

2. The source side of the parallel corpus (using the available supervised POS tagger) and common vector representation of words are combined to train the RNN (Algorithm 2).

3. The RNN trained is adapted in a light supervision manner, using a small monolingual target corpus (manually annotated) and the common vector representation of words (extracted from the initial parallel corpus).

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 training language (light supervision).

5 Experiments and Results

5.1 Data and tools

Initially, we applied our method to the English–French language pair. French was considered as the target language here. French is certainly not a resource-poor language, but it was used as if no tagger was available (in fact, TreeTagger (Schmid, 1995), a supervised POS tagger exists for this language and helps us to obtain a ground truth for

Lang. / Model	French		German		Greek		Spanish	
	All words	OOV	All words	OOV	All words	OOV	All words	OOV
Simple Projection	80.3%	77.1%	78.9%	73%	77.5%	72.8%	80%	79.7%
RNN-640-160	78.5%	70%	76.1%	76.4%	75.7%	70.7%	78.8%	72.6%
Projection+RNN	**84.5%**	**78.8%**	81.5%	**77%**	78.3%	**74.6%**	83.6%	**81.2%**
(Das, 2011)	—	—	82.8%	—	**82.5%**	—	**84.2%**	—
(Duong, 2013)	—	—	**85.4%**	—	80.4%	—	83.3%	—
(Gouws, 2015a)	—	—	84.8%	—	—	—	82.6%	—

Table 1: Unsupervised model : token-level POS tagging accuracy for Simple Projection, RNN [4], Projection+RNN and methods of Das & Petrov (2011), Duong et al (2013) and Gouws & Søgaard (2015).

evaluation). To train the RNN POS tagger, we used a training set of $10,000$ parallel sentences extracted from the ARCADE II English–French corpus (Veronis et al., 2008). Our validation corpus contains 1000 English sentences (these sentences are not in the train set) extracted from the ARCADE II English corpus. The test corpus is also extracted from the ARCADE II corpus, and it contains 1000 French sentences (which are obviously different from the train set) tagged with the French *TreeTagger* Toolkit (Schmid, 1995) and manually checked.

Encouraged by the results obtained on the English–French language pair, and in order to confirm our results, we run additional experiments on other languages, we applied our method to build RNN POS taggers for three more target languages — German, Greek and Spanish — with English as the source language, in order to compare our results with those of (Das and Petrov, 2011; Duong et al., 2013; Gouws and Søgaard, 2015a). Our training and validation (English) data extracted from the Europarl corpus (Koehn, 2005) are a subset of the training data of (Das and Petrov, 2011; Duong et al., 2013). The sizes of the data sets are: $65,000$ (train) and $10,000$ (dev) bi-sentences. For testing, we used the same test corpora (from CoNLL shared tasks on dependency parsing (Buchholz and Marsi, 2006)) as (Das and Petrov, 2011; Duong et al., 2013; Gouws and Søgaard, 2015a). The evaluation metric (*per-token* accuracy) and the Universal Tagset are the same as before. The source sides of the training corpora (ARCADE II and Europarl) and the validation corpora are tagged with the English *TreeTagger* Toolkit. Using the matching provided by Petrov et al. (2012) we map the TreeTagger and the CoNLL tagsets to a common Universal Tagset.

In order to build our unsupervised tagger based on a Simple Cross-lingual Projection (Algorithm 1), we tag the target side of the training corpus, with tags projected from English side through word-alignments established by GIZA++. After tags projection we use TNT Tagger to induce a target language POS Tagger (see Algorithm 1 described in Section 3.1).

Also, our proposed approach implements Algorithm 2 described before. We had to slightly modify the Recurrent Neural Network Language Modeling Toolkit (RNNLM) provided by Mikolov et al. (2011), to learn our Recurrent Neural Network Based POS Tagger[5]. The modifications include: (1) building the cross-lingual word representations automatically; and (2) learning and testing models with several hidden layers (common representation as input and universal POS tags as output).

The combined model is built for each considered language using cross-validation on the test corpus. First the test corpus is split into 2 equal parts and on each part, we estimate the interpolation parameter μ (Equation 1) which maximizes the *per-token* accuracy score. Then each part of test corpus is tagged using the combined model tuned (Equation 2) on the other part, and vice versa (standard cross-validation procedure).

Finally, we investigate how the performance of the adapted model changes according to target adaptation corpus size. We choose German as target adaptation language, because we dispose of a large German annotated data set (from CoNLL shared

[4]For RNN a single system is used for German, Greek and Spanish

[5]The modified source code is Available from the following URL `https://github.com/othman-zennaki/RNN_POS_Tagger.git`

tasks on dependency parsing). Then, we generate German adaptation sets of 7 different sizes (from 100 to 10,000 utterances). Each adaptation set is used to adapt our unsupervised RNN POS tagger. As contrastive experiments, we also learn supervised POS Taggers based on RNN, TNT or their linear combination.

5.2 Results and discussion

5.2.1 Unsupervised model

In table 1 we report the results obtained for the unsupervised approach. Preliminary RNN experiments used one hidden layer, but we obtained lower performance compared to those with two hidden layers. So we report here RNN accuracy achieved using two hidden layers, containing respectively 640 and 160 neurons (RNN-640-160). As shown in the table, this accuracy is close to that of the simple projection tagger, the difference coming mostly from out-of-vocabulary (OOV) words. As OOV words are not in the training corpus, their vector representations are empty (they contain only 0), therefore the RNN model uses only the context information, which is insufficient to tag correctly the OOV words in the test corpus. We also observe that both methods seem complementary since the best results are achieved using the linearly combined model Projection+RNN-640-160. It achieves comparable results to Das and Petrov (2011), Duong et al. (2013) (who used the full Europarl corpus while we used only a 65,000 subset of it) and Gouws and Søgaard (2015a) (who in addition used Wiktionary and Wikipedia) methods. It is also important to note that a single RNN tagger applies to German, Greek and Spanish; so this is a truly multilingual POS tagger! Therefore, as for several other NLP tasks such as language modelling or machine translation (where standard and NN-based models are combined in a log-linear model), the use of both standard and RNN-based approaches seems necessary to obtain optimal performances.

In order to know in what respect using RNN improves combined model accuracy, and vice versa, we analyzed the French test corpus. In the example provided in table 2, RNN information helps to resolve the French word *"précise"* tag ambiguity: in the Simple Projection model it is tagged as a verb (VERB), whereas it is an adjective (ADJ) in this particular context. We hypothesize that the context information is better represented in RNN, because of the recurrent connections.

English	a precise breakdown of spending
French	une répartition précise des dépenses
Simple Projection	une/DET répartition/NOUN précise/**VERB** des/ADP ...
Projection + RNN	une/DET répartition/NOUN précise/**ADJ** des/ADP ...

Table 2: Improved tagged example for french target language.

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).

EN Supervised Treetagger	... other/ADJ specific/ADJ groups/NOUN ...
FR Unsupervised RNN	... autres/ADJ groupes/NOUN spcifiques/ADJ ...

Table 3: Word order divergence -unambiguous tag word-.

EN Supervised Treetagger	... two/NUM local/ADJ groups/NOUN ...
FR Unsupervised RNN	... deux/NUM groupes/NOUN locaux/**NOUN** ...
Projection + RNN	... deux/NUM groupes/NOUN locaux/ **ADJ** ...

Table 4: Word order divergence -ambiguous tag word-.

5.2.2 Lightly supervised model

In table 5 we report the results obtained after adaptation with a gradually increasing amount of

DE Corpus Size / Model	0	100	500	1k	2k	5k	7k	10k
Unsupervised RNN + DE Adaptation	76.1%	**82.1%**	**87.3%**	**90.4%**	90.7%	91.2%	91.4%	92.4%
Supervised RNN DE only	—	71%	76.4%	82.1%	90.6%	93%	94.2%	95.2%
Supervised TNT DE only	—	80.5%	86.5%	89%	92.2%	94.1%	95.3%	95.7%
Supervised RNN + Supervised TNT DE	—	81%	86.7%	90.1%	**94.2%**	**95.3%**	**95.7%**	**96%**

Table 5: Lightly supervised model : effect of German adaptation corpus (manually annotated) size on method described in Section 4 (Unsupervised RNN + DE Adaptation trained on EN Europarl and adapted to German). Contrastive experiments with German supervised POS taggers using same data (RNN, TNT and RNN+TNT). 0 means no German corpus used during training.

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 $V_{wi}, i = 1, ..., N$, 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 (< 1000 target language utterances). When more data is available (> 1000 utterances), the supervised approaches start to be better (but RNN and TNT are still complementary since their combination improves the tag accuracy).

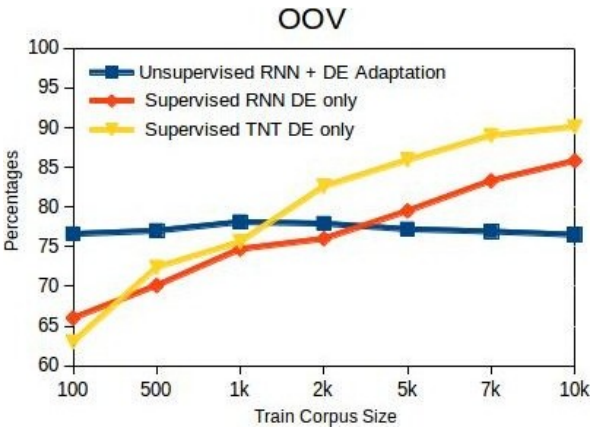

Figure 2: Accuracy on OOV according to German training corpus size for Unsupervised RNN + DE Adaptation, Supervised RNN DE and Supervised TNT DE.

Figure 2 details the behavior of the same methods for OOV words. We clearly see the limitation of the Unsupervised RNN + Adaptation to handle OOV words, 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. Better handling OOV words in unsupervised RNN training is our priority for future works.

Finally, these results show that for all training data sizes, RNN brings complementary information on top of a more classical approach such as TNT.

6 Conclusion

In this paper, we have presented a novel approach which uses a language-independent word representation (based only on word occurrence in a parallel corpus) within a recurrent neural network (RNN) to build multilingual POS tagger. Our method induces automatically POS tags from one language to another (or several others) and needs only a parallel corpus and a POS tagger in the source language (without using word alignment information).

We first empirically evaluated the proposed approach on two unsupervised POS taggers based on RNN : (1) English–French cross-lingual 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 ($65,000$) of Europarl corpus used.

Additionally, when a small amount of supervised data is available, the experimental results demonstrated the effectiveness of our method in a weakly supervised context (especially for very-low-resourced settings).

Although our initial experiments are positive, we believe they can be improved in a number of ways. In future work, we plan, on the one hand, to better manage OOV representation (for instance using Cross-lingual Word Embeddings), and, on the other hand, to consider more complex tasks such as word senses projection or semantic role labels projection.

References

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

P. Annesi and R. Basili. 2010. Cross-lingual alignment of FrameNet annotations through Hidden Markov Models, In *Proceedings of CICLing* :12–25.

Y. Bengio, H. Schwenk, J. Senécal, F. Morin and J. Gauvain. 2006. Neural probabilistic language models, In *Innovations in Machine Learning*:137–186.

L. Bentivogli, P. Forner and E. Pianta. 2004. Evaluating cross-language annotation transfer in the MultiSemCor corpus, In *Proceedings of the 20th international conference on Computational Linguistics*:364–370. Association for Computational Linguistics.

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

T. Brants. 2000. TnT: a statistical part-of-speech tagger, In *Proceedings of the sixth conference on Applied natural language processing*:224–231.

R. Collobert and J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning, In *Proceedings of the International Conference on Machine Learning (ICML)*:160–167.

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

D. Das and S. Petrov. 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*, volume 1:600–609. Association for Computational Linguistics.

L. Duong, P.Cook, S. Bird and P. Pecina. 2013. Simpler unsupervised POS tagging with bilingual projections, In *ACL (2)* :634–639.

G. Durrett, A. Pauls and D. Klein. 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*:1–11. Association for Computational Linguistics.

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

J. Henderson. 2004. Discriminative training of a neural network statistical parser, In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*:95–102.

S. Federici and V. Pirrelli. 1993. Analogical modelling of text tagging, *unpublished report*, Istituto diLinguistica Computazionale, Pisa, Italy.

S. Gouws and A. Søgaard 2015. Simple task-specific bilingual word embeddings, In *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL'15*:1386–1390.

S. Gouws, Y. Bengio and G. Corrado 2015. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments, In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*:748–756.

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

S. Kim, K. Toutanova and H. Yu, 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: Long Papers-volume 1*:694–702. Association for Computational Linguistics.

S. Li, J.V Graça and B. Taskar. 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*:1389–1398. Association for Computational Linguistics.

T. Mikolov, M. Karafiát , L. Burget, J. Cernocký and S. Khudanpur. 2010. Recurrent neural network based language model, In *INTERSPEECH*:1045–1048.

T. Mikolov, S. Kombrink , A. Deoras, L. Burget and J. Cernocky. 2011. RNNLM-Recurrent neural network language modeling toolkit, In *Proc. of the 2011 ASRU Workshop*:196–201.

F.Och and H.Ney. 2000. Improved Statistical Alignment Models, In *ACL00*:440–447.

S. Padó. 2007. Cross-Lingual Annotation Projection Models for Role-Semantic Information, In *German Research Center for Artificial Intelligence and Saarland University*, volume 21.

S. Petrov, D. Das and R. McDonald. 2012. A Universal Part-of-Speech Tagset, In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC '12)*:2089–2096.

P. Koehn 2005. Europarl: A parallel corpus for statistical machine translation, In *MT summit*, volume 5 :79–86.

D. Rumelhart, E. Hinton and R.J. Williams . 1985. Learning internal representations by error propagation,

In *Learning internal representations by error propagation* .

H. Schmid. 1995. TreeTagger— a Language Independent Part-of-speech Tagger, In *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, volume 43 :28.

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

O. Täckström, R. McDonald and J. Uszkoreit. 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*:477–487. Association for Computational Linguistics.

O. Täckström, R. McDonald, J. Nivre. 2013. Target language adaptation of discriminative transfer parsers, In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech-nologies (NAACL)*.

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

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

L. Van der Plas and M. Apidianaki. 2014. Cross-lingual Word Sense Disambiguation for Predicate Labelling of French, In *Proceedings of the 21st TALN (Traitement Automatique des Langues Naturelles) conference* :46–55.

J. Veronis, O. Hamon, C. Ayache, R. Belmouhoub, O. Kraif, D. Laurent, T.M.H. Nguyen, N. Semmar, F. Stuck and Z. Wajdi. 2008. Arcade II Action de recherche concerte sur l'alignement de documents et son valuation, Chapitre 2, *Editions Hermés* .

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

G. Wisniewski, N. Pécheux, S. Gahbiche-Braham and F. Yvon. 2014. Cross-Lingual Part-of-Speech Tagging through Ambiguous Learning, In *EMNLP'14*:1779–1785.

M. Xiao and Y. Guo. 2014. Distributed Word Representation Learning for Cross-Lingual Dependency Parsing, In *CoNLL-2014*:119–129.

D. Yarowsky, G. NGAI and R. Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora, In *Proceedings of the first international conference on Human language technology research*:1–8. Association for Computational Linguistics.