Inducing Multilingual Text Analysis Tools Using Bidirectional Recurrent Neural Networks

Othman Zennaki, Nasredine Semmar, Laurent Besacier

Inducing Multilingual Text Analysis Tools Using Bidirectional Recurrent Neural Networks

Othman Zennaki
CEA, LIST, LVIC
Gif-sur-Yvette, France
othman.zennaki@cea.fr

Nasredine Semmar
CEA, LIST, LVIC
Gif-sur-Yvette, France
nasredine.semmar@cea.fr

Laurent Besacier
LIG, Univ. Grenoble-Alpes
Grenoble, France
laurent.besacier@imag.fr

Abstract

This work focuses on the rapid development of linguistic annotation tools for resource-poor languages. We experiment 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 the source and target language. More precisely, our method has the following characteristics: (a) it does not use word alignment information, (b) it does not assume any knowledge about foreign languages, which makes it applicable to a wide range of resource-poor languages, (c) it provides truly multilingual taggers. We investigate both uni- and bi-directional RNN models and propose a method to include external information (for instance low level information from POS) 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 POS and super sense taggers.

1 Introduction

In order to minimize the need for annotated resources (produced through manual annotation, or by manual check of automatic annotation), several research works were interested in building Natural Language Processing (NLP) tools based on unsupervised or semi-supervised approaches (Collins and Singer, 1999; Klein, 2005; Goldberg, 2010). For example, NLP tools based on cross-language projection of linguistic annotations achieved good performances in the early 2000s (Yarowsky et al., 2001). The key idea of annotation projection can be summarized as follows: through word alignment in parallel text corpora, the annotations are transferred from the *source* (resource-rich) language to the *target* (under-resourced) language, and the resulting annotations are used for supervised training in the target language. However, automatic word alignment errors (Fraser and Marcu, 2007) limit the performance of these approaches.

Our work is built upon these previous contributions and observations. We explore the possibility of using Recurrent Neural Networks (RNN) to build multilingual NLP tools for resource-poor languages analysis. The major difference with previous works is that we do not explicitly use word alignment information. Our only assumption is that parallel sentences (source-target) are available and that the source part is annotated. In other words, we try to infer annotations in the target language from sentence-based alignments only. While most NLP researches on RNN have focused on monolingual tasks[1] and sequence labeling (Collobert et al., 2011; Graves, 2012), this paper, however, considers the problem of learning multilingual NLP tools using RNN.

Contributions In this paper, we investigate the effectiveness of RNN architectures — Simple RNN (SRNN) and Bidirectional RNN (BRNN) — for multilingual sequence labeling tasks without using any word alignment information. Two NLP tasks are considered: Part Of Speech (POS) tagging and Super Sense (SST) tagging (Ciaramita and Altun, 2006). Our RNN architectures demonstrate very competitive results on unsupervised training for new target languages. In addition, we show that the integration of

[1]Exceptions are the recent propositions on Neural Machine Translation (Cho et al., 2014; Sutskever et al., 2014)

POS information in RNN models is useful to build multilingual coarse-grain semantic (Super Senses) taggers. For this, a simple and efficient way to take into account low-level linguistic information for more complex sequence labeling RNN is proposed.

Methodology For training our multilingual RNN models, we just need as input a parallel (or multi-parallel) corpus between a resource-rich language and one or many under-resourced languages. Such a parallel corpus can be manually obtained (clean corpus) or automatically obtained (noisy corpus).

To show the potential of our approach, we investigate two sequence labeling tasks: cross-language POS tagging and multilingual Super Sense Tagging (SST). For the SST task, we measure the impact of the parallel corpus quality with manual or automatic translations of the SemCor (Miller et al., 1993) translated from English into Italian (manually and automatically) and French (automatically).

Outline The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes our cross-language annotation projection approaches based on RNN. Section 4 presents the empirical study and associated results. We finally conclude the paper in Section 5.

2 Related Work

Cross-lingual projection of linguistic annotations was pioneered by Yarowsky et al. (2001) who created new monolingual resources by transferring annotations from resource-rich languages onto resource-poor languages through the use of word alignments. The resulting (noisy) annotations are used in conjunction with robust learning algorithms to build cheap unsupervised NLP tools (Padó and Lapata, 2009). This approach has been successfully used to transfer several linguistic annotations between languages (efficient learning of POS taggers (Das and Petrov, 2011; Duong et al., 2013) and accurate projection of word senses (Bentivogli et al., 2004)). 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 (Och and Ney, 2000). However, the noisy (non perfect) outputs of these methods is a serious limitation for the annotation projection based on word alignments (Fraser and Marcu, 2007).

To deal with this limitation, 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 et al., 2012; Al-Rfou et al., 2013; Täckström et al., 2013; Luong et al., 2015; Gouws and Søgaard, 2015; Gouws et al., 2015). 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. 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 et al., 2012), cross-lingual document classification and lexical translation task (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 next section, is inspired by these works since we also try to induce a common language-independent feature space (crosslingual 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 the source and target languages[2] our common multilingual representation is very agnostic. We use a simple (multilingual) vector representation based on the occurrence 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 or POS tagging).

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 simple and bidirectional RNN architectures (SRNN and BRNN respectively). 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 multilingual SST.

[2] to train a bilingual representation regardless of the task

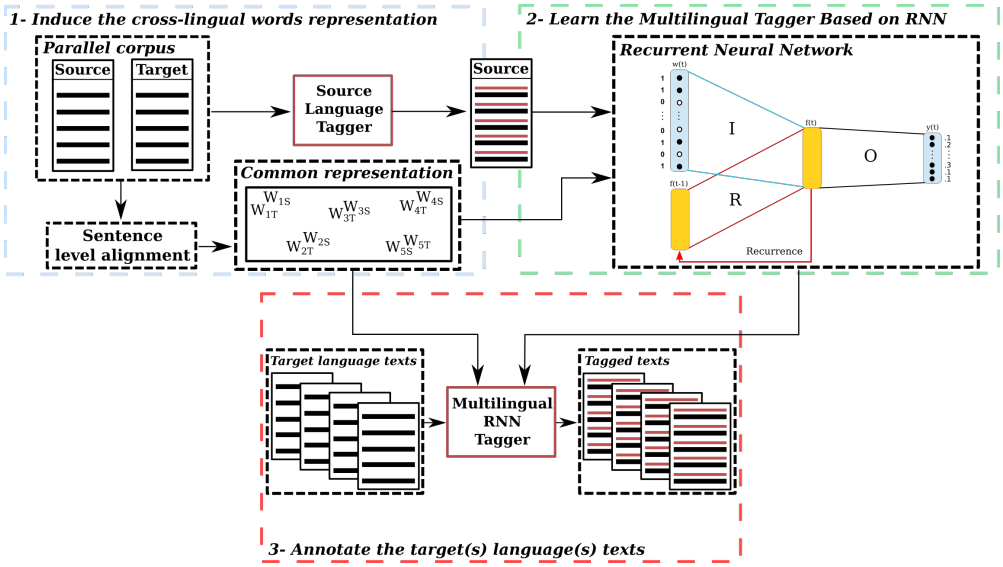

Figure 1: Overview of the proposed model architecture for inducing multilingual RNN taggers.

3 Unsupervised Approach Overview

To avoid projecting label information from deterministic and error-prone word alignments, we propose to represent the word alignment information intrinsically in a recurrent neural network architecture. The idea consists in implementing a recurrent neural network as a multilingual sequence labeling tool (we investigate POS tagging and SST tagging). Before describing our cross-lingual (multilingual if a multi-parallel corpus is used) neural network tagger, we present the simple cross-lingual projection method, considered as our baseline in this work.

3.1 Baseline Cross-lingual Annotation Projection

We use direct transfer as a baseline system which is similar to the method described in (Yarowsky et al., 2001). First we tag the source side of the parallel corpus using the available supervised tagger. Next, we align words in the parallel corpus to find out corresponding source and target words. Tags are then projected to the (resource-poor) target language. The target language tagger is trained using any machine learning approach (we use TNT tagger (Brants, 2000) in our experiments).

3.2 Proposed Approach

We propose a method for learning multilingual sequence labeling tools based on RNN, as it can be seen in Figure 1. In our approach, a parallel or multi-parallel corpus between a resource-rich language and one or many under-resourced languages is used to extract common (multilingual) and agnostic words representations. 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 truly multilingual and can be also used to tag texts in target language(s).

3.2.1 Common Words Representation

In our *agnostic* representation, we associate to each word (in source *and* 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 $V_{wi} = 1$.

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 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 and it lets the RNN learns the optimal internal representation needed for the annotation task (for instance, the hidden layers of the RNN can be considered as multi-lingual embeddings of the words).

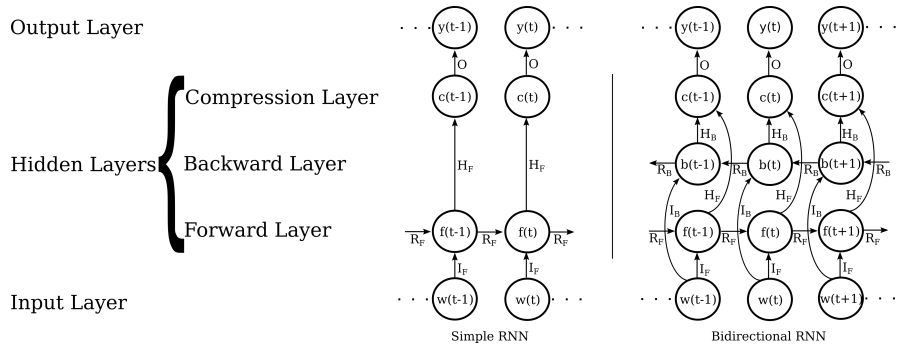

Figure 2: High level schema of RNN used in our work.

3.2.2 Recurrent Neural Networks

There are two major architectures of neural networks: Feedforward (Bengio et al., 2003) and Recurrent Neural Networks (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 recurrent neural networks 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): *Simple* RNN (SRNN) and *Bidirectional* RNN (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 (POS) information for multilingual Super Sense Tagging (SST).

A. Simple RNN

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

The architecture of the SRNN considered in this work is shown in Figure 2. In this architecture, we have 4 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 bi-sentences 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 80-1024 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 :

$$f(t) = \Sigma(w(t).I_F(t) + f(t-1).R_F(t)) \quad (1)$$

$$c(t) = \Sigma(f(t).H_F(t)) \quad (2)$$

$$y(t) = \Gamma(c(t).O(t)) \quad (3)$$

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.

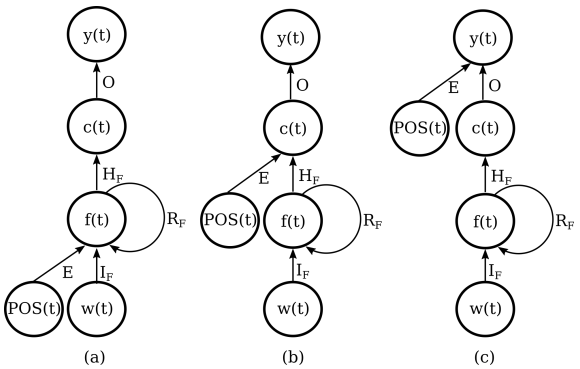

Figure 3: SRNN variants with POS information at three levels: (a) input layer, (b) forward layer, (c) compression layer.

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 SRNN 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 SRNN 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 SRNN, a simple extension of the SRNN architecture — namely Bidirectional recurrent neural network (BRNN) (Schuster and Paliwal, 1997) — is used to ensure that context at previous and future time steps will be considered.

B. 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 SRNN.

C. RNN Variants

As mentioned in the introduction, 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 this low level information should be included to improve SST performance? As specified in Figure 3, 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).

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 et al., 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 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 cross-lingual projection model in our experiments.

3.3 Dealing with out-of-vocabulary words

For the words absent from in the initial parallel corpus, their vector representation is a vector of zero values. Consequently, during testing, the RNN 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[3], 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. 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 simple cross-lingual projection 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

$$P_{M12}(t|w) = (\mu P_{M1}(t|w, C_{M1}) + (1 - \mu)P_{M2}(t|w, C_{M2})) \quad (4)$$

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 the function f described as

$$f(w) = \arg\max_t(P_{M12}(t|w)) \quad (5)$$

4 Experiments

Our models are evaluated on two labeling tasks: Cross-language Part-Of-speech (POS) tagging and Multilingual Super Sense Tagging (SST).

4.1 Multilingual POS Tagging

We applied our method to build RNN POS taggers for four target languages - French, German, Greek and Spanish - with English as the source language.

In order to determine the effectiveness of our common words representation described in section 3.2.1, we also investigated the use of state-of-the-art bilingual word embeddings (using MultiVec Toolkit (Bérard et al., 2016)) as input to our RNN.

4.1.1 Dataset

For French as a target language, we used a training set of $10,000$ parallel sentences, a validation set of 1000 English sentences, and a test set of 1000 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; Duong et al., 2013). This choice allows us to compare our results with those of (Das and Petrov, 2011; Duong et al., 2013; Gouws and Søgaard, 2015). The train data set contains $65,000$ bi-sentences ; a validation set of $10,000$ bi-sentences is also available. For testing, we use the same test corpora as (Das and Petrov, 2011; Duong et al., 2013; Gouws and Søgaard, 2015) (bi-sentences from CoNLL shared

[3] words which do not have a known vector representation

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	73.0	77.5	72.8	80.0	79.7
SRNN MultiVec	75.0	65.4	70.3	68.8	71.1	65.4	73.4	62.4
SRNN	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
Projection + SRNN	84.5	78.8	81.5	77.0	78.3	74.6	83.6	81.2
Projection + BRNN	85.2	79.0	81.9	77.1	79.2	75.0	84.4	81.7
Projection + BRNN - OOV	**85.6**	**80.4**	82.1	**78.7**	79.9	**78.5**	**84.4**	**81.9**
(Das, 2011)	—	—	82.8	—	**82.5**	—	84.2	—
(Duong, 2013)	—	—	**85.4**	—	80.4	—	83.3	—
(Gouws, 2015a)	—	—	84.8	—	—	—	82.6	—

Table 1: Token-level POS tagging accuracy for Simple Projection, SRNN using MultiVec bilingual word embeddings as input, RNN[5], Projection+RNN and methods of Das & Petrov (2011), Duong et al (2013) and Gouws & Søgaard (2015).

tasks on dependency parsing (Buchholz and Marsi, 2006)). The evaluation metric (*per-token* accuracy) and the Petrov et al. (2012) *universal tagset* are used for evaluation.

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 the TreeTagger and the CoNLL tagsets to the common *Universal Tagset*.

In order to build our baseline unsupervised tagger (based on a Simple Cross-lingual Projection – see section 3.1), we also tag the target side of the training corpus, with tags projected from English side through word-alignments established by GIZA++. After 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 2 equal parts and on each part, we estimate the interpolation parameter μ (Equation 4) which maximizes the *per-token* accuracy score. Then each part of test corpus is tagged using the 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.

4.1.2 Results and discussion

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

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), 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 3.2.1) seem to perform significantly better than RNN models using MultiVec bilingual word embeddings.

It is also important to note that only one single SRNN and BRNN tagger applies to German, Greek and Spanish; so this is a truly multilingual POS tagger! Finally, as for several other NLP tasks such as language modelling or machine translation (where standard and NN-based models are generally combined in order to obtain optimal results), the combination of standard and RNN-based approaches (*Projection+_*) seems necessary to further optimize POS tagging accuracies.

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

4.2 Multilingual SST

In order to measure the impact of the parallel corpus quality on our method, we also learn our 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.

4.2.1 Dataset

SemCor The SemCor (Miller et al., 1993) is a subset of the Brown Corpus (Kucera and Francis, 1979) labeled with the *WordNet* (Fellbaum, 1998) senses.

MultiSemCor The English-Italian MultiSemcor (MSC-IT-1) corpus is a manual translation of the English SemCor to Italian (Bentivogli et al., 2004). As we already mentioned, we are also interested in measuring the impact of the 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 system (Besacier et al., 2012) to translate the English SemCor to French (MSC-FR-1).

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 learn 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 Word Sense Disambiguation) (Navigli et al., 2013) test corpora, which are available in 5 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.

4.2.2 SST Systems Evaluated

The goals of our SST experiments are twofold: first, to investigate the effectiveness of using POS information to build multilingual super sense tagger, secondly to measure the impact of the parallel corpus quality (manual or automatic translation) on our RNN models (SRNN, BRNN and our proposed variants). To summarize, we build four super sense taggers based on baseline cross-lingual projection (see section 3.1) using four versions of MultiSemcor (MSC-IT-1, MSC-IT-2, MSC-FR-1, MSC-FR-2) described above. Then we use the same four versions to train our multilingual SST models based on SRNN and BRNN. For learning our multilingual SST models based on RNN variants proposed in part (C) of section 3.2.2, we also tag SemCor using *TreeTagger* (POS tagger proposed by Schmid (1995)).

4.2.3 Results and discussion

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

The DAEBAK! (Navigli and Lapata, 2010) and the UMCC-DLSI systems (Gutiérrez Vázquez et al., 2011) have also participated to SemEval 2013 Task 12. However, they use a supervised approach [6].

Table 2 shows the results obtained by our RNN models and by two SemEval 2013 WSD systems. SRNN-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 combination of simple projection and RNN which confirms (as for POS tagging) that both approaches are complementary.

[6] DAEBAK! and UMCC-DLSI for SST have obtained: 68.1% and 72.5% on Italian; 59.8% and 67.6 % on French

	Model	Italian		French	
		MSC-IT-1 trans man.	**MSC-IT-2 trans. auto**	**MSC-FR-1 trans. auto**	**MSC-FR-2 trans auto.**
Baseline	Simple Projection	61.3	45.6	42.6	44.5
SST Based RNN	SRNN	59.4	46.2	46.2	47.0
	BRNN	59.7	46.2	46.0	47.2
	SRNN-POS-In	61.0	47.0	46.5	47.3
	SRNN-POS-H1	59.8	46.5	46.8	47.4
	SRNN-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	Projection + SRNN	62.0	46.7	46.5	47.4
	Projection + BRNN	62.2	46.8	46.4	47.5
	Projection + SRNN-POS-In	62.9	47.4	46.9	47.7
	Projection + SRNN-POS-H1	62.5	47.0	47.1	48.0
	Projection + SRNN-POS-H2	63.5	49.2	48.0	50.1
	Projection + BRNN-POS-In	62.9	47.5	46.9	47.8
	Projection + BRNN-POS-H1	62.7	47.0	47.0	48.0
	Projection + BRNN-POS-H2	63.6	49.3	48.0	50.3
	Projection + 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	

Table 2: Super Sense Tagging (SST) accuracy for Simple Projection, RNN and their combination.

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.

5 Conclusion

In this paper, we have presented an approach based on recurrent neural networks (RNN) to induce multilingual text analysis tools. We have studied Simple and Bidirectional RNN architectures on multilingual POS and SST tagging. We have also proposed new RNN variants in order to take into account low level (POS) information in a super sense tagging task. Our approach 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 truly multilingual taggers (1 tagger for N languages) (c) it can be easily adapted to a new target language (when a small amount of supervised data is available, a previous study (Zennaki et al., 2015a; Zennaki et al., 2015b) has shown the effectiveness of our method in a weakly supervised context).

Short term perspectives are to apply multi-task learning to build systems that simultaneously perform syntactic and semantic analysis. Adding out-of-language data to improve our RNN taggers is also possible (and interesting to experiment) with our common (multilingual) vector representation.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. *CoNLL-2013*, pages 183–192.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Luisa Bentivogli, Pamela Forner, and Emanuele Pianta. 2004. Evaluating cross-language annotation transfer in the multisemcor corpus. In *COLING*, page 364.

Alexandre Bérard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. 2016. Multivec: a multilingual and multilevel representation learning toolkit for nlp. In *The 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*.

Laurent Besacier, Benjamin Lecouteux, Marwen Azouzi, and Ngoc-Quang Luong. 2012. The lig english to french machine translation system for iwslt 2012. In *IWSLT*, pages 102–108.

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

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth CoNLL*, pages 149–164.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. *Syntax, Semantics and Structure in Statistical Translation*.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the EMNLP-2006*, pages 594–602.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of the joint SIGDAT conference on EMNLP and very large corpora*, pages 100–110.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. *Proceedings of the 49th ACL*, 1:600–609.

Long Duong, Paul Cook, Steven Bird, and Pavel Pecina. 2013. Simpler unsupervised pos tagging with bilingual projections. In *ACL (2)*, pages 634–639.

Greg Durrett, Adam Pauls, and Dan Klein. 2012. Syntactic transfer using a bilingual lexicon. In *The Joint Conference on EMNLP and CoNLL*, pages 1–11.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Andrew Brian Goldberg. 2010. *New directions in semi-supervised learning*. Ph.D. thesis, University of Wisconsin–Madison.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *NAACL-HLT*, pages 1386–1390.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. *ICML 2015*.

Alex Graves. 2012. *Supervised sequence labelling*. Springer.

Yoan Gutiérrez Vázquez, Antonio Fernández Orquín, Andrés Montoyo Guijarro, Sonia Vázquez Pérez, et al. 2011. Enriching the integration of semantic resources based on wordnet.

Dan Klein. 2005. *The unsupervised learning of natural language structure*. Ph.D. thesis, Stanford University.

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

H Kucera and W Francis. 1979. A standard corpus of present-day edited american english, for use with digital computers (revised and amplified from 1967 version).

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

Tomáš Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010*, pages 1045–1048.

Tomáš Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on HLT*, pages 303–308.

Roberto Navigli and Mirella Lapata. 2010. An experimental study of graph connectivity for unsupervised word sense disambiguation. *Pattern Analysis and Machine Intelligence*, 32(4):678–692.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 : Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics*, volume 2, pages 222–231.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on ACL*, pages 440–447.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36(1):307–340.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *LREC'12*.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1985. Learning internal representations by error propagation. Technical report, DTIC Document.

Helmut Schmid. 1995. Treetagger| a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.

Jürgen Schmidhuber. 1992. A fixed size storage o (n3) time complexity learning algorithm for fully recurrent continually running networks. *Neural Computation*, 4(2):243–248.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *Signal Processing*, 45(11):2673–2681.

Didier Schwab, Jérôme Goulian, Andon Tchechmedjiev, and Hervé Blanchon. 2012. Ant colony algorithm for the unsupervised word sense disambiguation of texts: Comparison and evaluation. In *COLING*, pages 2389–2404.

Martin Sundermeyer, Ilya Oparin, J-L Gauvain, Ben Freiberg, Ralf Schluter, and Hermann Ney. 2013. Comparison of feedforward and recurrent neural network language models. In *ICASSP*, pages 8430–8434. IEEE.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 conference of the NAACL-HLT*, pages 477–487.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers.

Ivan Titov and Alexandre Klementiev. 2012. Crosslingual induction of semantic roles. In *Proceedings of the 50th Annual Meeting of the ACL*, volume 1, pages 647–656.

J Veronis, O Hamon, C Ayache, R Belmouhoub, O Kraif, D Laurent, TMH Nguyen, N Semmar, F Stuck, and W Zaghouani. 2008. Arcade ii action de recherche concertée sur l'alignement de documents et son évaluation.

David Yarowsky, Grace Ngai, and Richard 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*, pages 1–8.

Othman Zennaki, Nasredine Semmar, and Laurent Besacier. 2015a. Unsupervised and lightly supervised part-of-speech tagging using recurrent neural networks. In *PACLIC 29*.

Othman Zennaki, Nasredine Semmar, and Laurent Besacier. 2015b. Utilisation des réseaux de neurones récurrents pour la projection interlingue d'étiquettes morpho-syntaxiques à partir d'un corpus parallèle. In *TALN 2015*.