Deep Neural Networks for Syntactic Parsing of Morphologically Rich Languages

Joël Legrand[1,2] and **Ronan Collobert**[*3,1]

[1] Idiap Research Institute, Martigny, Switzerland

[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[3] Facebook AI Research, Menlo Park (CA), USA

Abstract

Morphologically rich languages (MRL) are languages in which much of the structural information is contained at the word-level, leading to high level word-form variation. Historically, syntactic parsing has been mainly tackled using generative models. These models assume input features to be conditionally independent, making difficult to incorporate arbitrary features. In this paper, we investigate the greedy discriminative parser described in (Legrand and Collobert, 2015), which relies on word embeddings, in the context of MRL. We propose to learn morphological embeddings and propagate morphological information through the tree using a recursive composition procedure. Experiments show that such embeddings can dramatically improve the average performance on different languages. Moreover, it yields state-of-the art performance for a majority of languages.

1 Introduction

Morphologically rich languages (MRL) are languages for which important information concerning the syntactic structure is expressed through word formation, rather than constituent-order patterns. Unlike English, they can have complex word structure as well as flexible word order. A common practice when dealing with such languages is to incorporate morphological information explicitly (Tsarfaty et al., 2013). However this poses two problems to the classical generative models: they assume input features to be conditionally independent which makes the incorporation of arbitrary features difficult. Moreover, refining input features leads to a data sparsity issue.

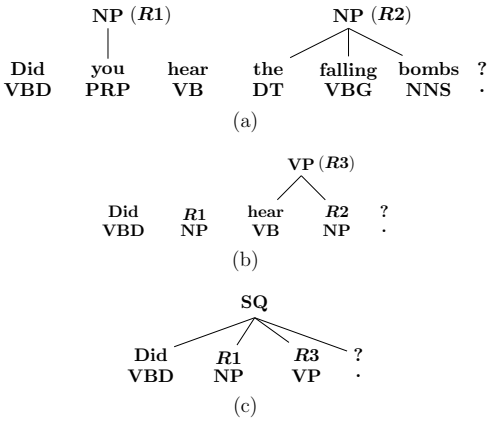
In the other hand, neural network-based models using continuous word representations as input have been able to overcome the data sparsity problem inherent in NLP (Huang and Yates, 2009). Furthermore, neural networks allow to incorporate arbitrary features and learn complex non-linear relations between them. Legrand and Collobert (2015) introduced a greedy syntactic parser, based on neural networks which relies on word embeddings. This model maintains a history of the previous node predictions, in the form of vector representations, by leveraging a recursive composition procedure.

In this paper, we propose to enhance this model for syntactic parsing of MRL, by learning morphological embeddings. We take advantage of a recursive composition procedure similar to the one used in (Legrand and Collobert, 2015) to propagate morphological information during the parsing process. We evaluate our approach on the SPMRL (Syntactic Parsing of MRL) Shared Task 2014 (Seddah et al., 2013) on nine different languages. Each of them comes with a set of morphological features allowing to augment words with information such as their grammatical functions, relation with other words in the sentence, prefixes, affixes and lemmas. We show that integrating morphological features allows to increase dramatically the average performance and yields state-of-the-art performance for a majority of languages.

1.1 Related work

Both the baseline (Berkeley parser) and the current state-of-the-art model on the SPMRL Shared Task 2014 (Björkelund et al., 2014) rely on probabilistic context free grammar (PCFG)-based features. The latter uses a product of PCFG with latent annotation based models (Petrov, 2010), with a coarse-to-

[*] All research was conducted at the Idiap Research Institute, before Ronan Collobert joined Facebook AI Research



	$\mathbf{I}_W$	:	Did	you	hear	the	falling	bombs ?
(a)	$\mathbf{I}_T$	:	VBD	PRP	VB	DT	VBG	NNS .
	O	:	O	S-NP	O	B-NP	I-NP	E-NP O
	$\mathbf{I}_W$	:	Did	*R1*	hear	*R2*	.	
(b)	$\mathbf{I}_T$	:	VBD	NP	VB	NP	.	
	O	:	O	O	B-VP	E-VP	.	
	$\mathbf{I}_W$	:	Did	*R1*	*R3*	?		
(c)	$\mathbf{I}_T$	:	VDB	NP	VP	.		
	O	:	B-SQ	I-SQ	I-SQ	E-SQ		

Figure 1: Greedy parsing algorithm (3 iterations), on the sentence "Did you hear the falling bombs ?". $\mathbf{I}_W$, $\mathbf{I}_T$ and **O** stand for input words (or composed word representations R_i), input syntactic tags (parsing or part-of-speech) and output tags (parsing), respectively. The tree produced after 3 greedy iterations can be reconstructed as the following: `(SQ (VBD Did) (NP (PRP you)) (VP (VB hear) (NP (DT the) (VBG falling) (NNS bombs))) (. ?))`.

fine decoding strategy. The output is then discriminatively reranked (Charniak and Johnson, 2005) to select the best analysis. In contrast, the parser used in this paper constructs the parse tree in a greedy manner and relies only on word, POS tags and morphological embeddings.

Several other papers have reported results for the SPMRL Shared Task 2014. (Hall et al., 2014) introduced an approach where, instead of propagating contextual information from the leaves of the tree to internal nodes in order to refine the grammar, the structural complexity of the grammar is minimized. This is done by moving as much context as possible onto local surface features. This work was refined in (Durrett and Klein, 2015), taking advantage of continuous word representations. The system used in this paper also leverages words embeddings but has two major differences. First, it proceeds step-by-step in a greedy manner where (Durrett and Klein, 2015) is using structured inference (CKY). Second, it leverages a compositional node feature which propagates information from the leaves to internal nodes, which is exactly what is claimed not to be done.

(Fernández-González and Martins, 2015) proposed a procedure to turn a dependency tree into a constituency tree. They showed that encoding order information in the dependency tree make it isomorphic to the constituent tree, allowing any dependency parser to produce constituents. Like the parser we used, their parser do not need to binarize the treebank as most of the others constituency parsers. Unlike this system, we do not use the dependency structure as an intermediate representation and directly perform constituency parsing over raw words.

2 Recurrent greedy parsing

In this paper, we used the model presented in (Legrand and Collobert, 2015). It is a NN-based model which performs parsing in a greedy recurrent way. It follows a bottom-up iterative procedure: the tree is built starting from the terminal nodes (sentence words), as shown in Figure 1. Each step can be seen as a sequence tagging task. A BIOES[1] prefixing scheme is used to rewrite this chunk (here node) prediction problem into a word tagging problem. Each iteration of the procedure merges input constituents into new nodes by applying the following steps:

- **Node tagger:** a neural network sliding window is applied over the input sequence of constituents (leaves or heads of trees predicted so far). This procedure (see Figure 2) outputs for each constituent a score s_i for each BIOES-prefixed parsing tag $t \in \mathcal{T}$ ($\mathcal{T}$ being the parsing tags ensemble).

- **Dynamic programming:** a coherent path of BIOES tags is retrieved by decoding over a constrained graph. This insures (for instance) that a B-A can be followed only by a I-A or a E-A (for all parsing tag A).

[1] (*B*egin, *I*ntermediate, *O*ther, *E*nd, *S*ingle)

- **Compositional procedure:** new nodes are created, merging input constituents, according to the dynamic programming predictions. A neural network composition module is then used to compute vector representations for the new nodes, according to the representations of the merged constituents, as well as their corresponding tags (POS or parsing).

The procedure ends when the top node is produced.

3 Parsing Morphologically Rich Languages

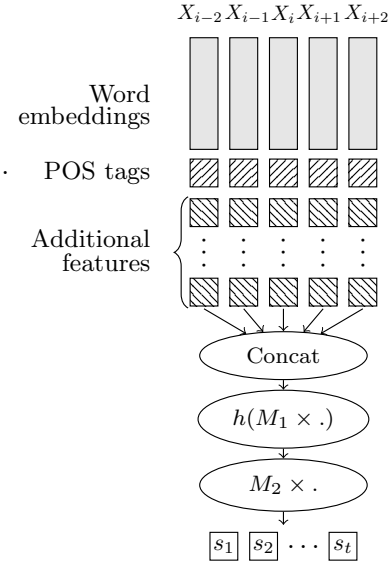


Figure 2: A constituent X_i (word or node previously predicted) is tagged by considering a fixed size context window of size K (here $K = 5$). The concatenated output of the compositional history and constituent tags is fed as input to the tagger. A standard two-layers neural network outputs a score s_i for each BIOES-prefixed parsing tag. Additional features can be easily fed to the network. Each category is assigned a new lookup table containing a vector of feature for every possible tag.

3.1 Morphological features

Morphological features enable the augmentation of input tokens with information expressed at a word level, such as grammatical function or relation to other words. For parsing MRL, they have proven to be very helpful (Cowan and Collins, 2005). The SMPRL corpus provides a different set of morphological features associated to the tree terminals (tokens) for every language. These features include morphosyntactic features such as case, number, gender, person and type, as well as specific morphological information such as verbal mood, proper/common noun distinction, lemma, grammatical function. They also include many language-specific features. For more details about the morphological features available, the reader can refer to (Seddah et al., 2013).

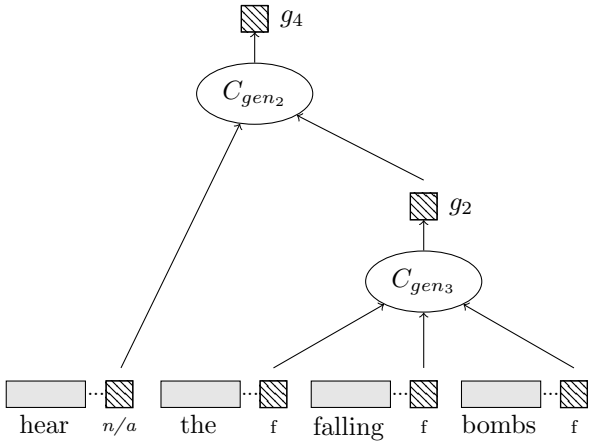


Figure 3: Recursive composition of the morphological feature *gender* (male (m) / female (f) / not applicable (n/a)). C_{gen_i} are the corresponding composition modules. The representation g_2 is first computed using the 3-inputs module C_{gen_3}. g_4 is obtained by using the 2-inputs module C_{gen_2}.

3.2 Morphological Embeddings

The parser from (Legrand and Collobert, 2015) relies only on word and tag embeddings. Besides these features, our model takes advantage of additional morphological features. As illustrated in Figure 2, each additional feature m is assigned a different lookup table containing morphological feature vectors of size d_m. The output vectors of the different morphological lookup-tables are simply concatenated to form the input of the next neural network layer.

3.3 Morphological composition

Morphological features are available only for leaves. To propagate morphological information to the nodes, we take advantage of a composition procedure similar to the one used in (Legrand and Collobert, 2015) for words and POS. As illustrated in Figure 3, every morphological feature m is assigned a set on composition modules C_{m_i} which take as input i morphological embeddings

Model	Ara.	Bas.	Fre.	Ger.	Heb.	Hun.	Kor.	Pol.	Swe.	AVG
Berkeley+POS	80.8	76.2	81.8	80.3	**92.2**	87.6	82.9	88.1	82.9	83.7
Berkeley RAW	79.1	69.8	80.4	79.0	87.3	81.4	73.3	79.5	78.9	78.7
(Björkelund et al., 2014)	82.2	90.0	84.0	82.1	91.6	**92.6**	86.5	88.6	**85.1**	87.0
Proposed approach	**84.1**	**91.0**	**85.7**	**84.6**	91.7	91.2	**87.8**	**94.1**	82.5	**88.1**

Table 1: Results for all languages in terms of F1-score, using gold POS and morphological tags. Berkeley+POS and Berkeley RAW are the two baseline system results provided by the organizers of the shared task. Our experiments used an ensemble of 5 models, trained starting from different random initializations.

of dimension d_m. Each composition module perform a matrix-vector operation followed by a non-linearity

$$C_{m_i}(x) = h(M_m^i.x)$$

where $M_m^i \in \mathbb{R}^{d_m \times id_m}$ is a matrix of parameters to be trained and h a pointwise non-linearity function. $x = [x_1...x_i]$ is the concatenation of the corresponding input morphological embeddings. Note that given a morphological feature we have a different matrix of weight for every possible size i. In practice most tree nodes do not merge more than a few constituents and we only consider composition sizes < 5.

4 Experiments

4.1 Corpus

Experiments were conducted on the SPMRL corpus provided for the Shared Task 2014 (Seddah et al., 2013). It provides sentences and tree annotations for 9 different languages (Arabic, Basque, French, German, Hebrew, Hungarian, Korean, Polish and Swedish) coming from various sources. For each language, *gold* part-of-speech and morphological tags are provided. Results for two baseline baseline system are provided in order to evaluate our models.

4.2 Setup

The model was trained using a stochastic gradient descent over the available training data. Hyper-parameters were tuned on the provided validation sets. The word embedding size and POS/parsing tag size were set to $D_{\mathcal{W}} = 100$ and $D_{\mathcal{T}} = 30$, respectively. The morphological tag embedding size was set to 10. The window size of the tagger was set to $K = 7$ and its number of hidden units to 300. All parameters were initialized randomly (including the words embeddings). As suggested in (Plaut and Hinton, 1987), the learning rate was divided by the size of the input vector of each layer. We applied the same dropout regularization as in (Legrand and Collobert, 2015).

4.3 Results

Table 2 presents the influence of adding morphological features to the model. We observe significant improvement for every languages except for Hebrew. On average, morphological features allowed to overcome the original model by 2 F1-score.

language	**Words + POS**	**+ morph**
Arabic	80.7	82.9
Basque	82.7	90.6
French	81.1	85.0
German	81.5	83.1
Hebrew	91.6	91.5
Hungarian	89.6	90.3
Korean	86.1	86.7
Polish	93.2	93.7
Swedish	81.1	81.5
AVG	85.3	87.3

Table 2: Influence of the additional morphological embeddings in terms of F1-score

Table 1 compares the performance in F1-score (obtained with the provided EVALB_SPMRL tool) of different systems, using the provided gold POS and morphological features. We compare our results with the two baselines provided with the task: (1) Berkeley parser with provided POS Tags (Berkeley+POS). (2) Berkeley Parser in raw mode where the parser do its own POS tagging (Berkeley RAW). We also report the results of the current state-of-the art model for this task (Björkelund et al., 2014). We included the same voting procedure as in citelegrand:2015, using 5 models trained starting from different random initializations. At

Model	Ara.	Bas.	Fre.	Ger.	Heb.	Hun.	Kor.	Pol.	Swe.	AVG
Berkeley+POS	78.7	74.7	79.8	78.3	85.4	85.2	78.6	86.7	80.6	80.9
Berkeley RAW	79.2	70.5	80.4	78.3	87.0	81.6	71.4	79.2	79.2	78.5
(Durrett and Klein, 2015)	80.2	85.4	81.2	80.9	88.6	90.7	82.2	93.0	83.4	85.1
(Fernández and Martins, 2015)	*n/a*	85.9	78.7	78.7	89.0	88.2	79.3	91.2	82.8	84.2
(Björkelund et al., 2014)	**81.3**	**87.9**	**81.8**	81.3	89.5	**91.8**	84.3	87.5	**84.0**	85.5
Proposed approach	80.4	87.5	80.8	**82.0**	**91.6**	90.0	**84.8**	**93.0**	80.5	**85.6**

Table 3: Results for all languages in terms of F1-score using predicted POS and morphological tags. Berkeley+POS and Berkeley RAW are the two baseline system results provided by the organizers of the shared task. Our experiments used an ensemble of 5 models, trained starting from different random initializations.

each iteration of the greedy parsing procedure, the BIOES-tag scores are averaged and the new node representations (words+POS and morphological composition) are computed for each model by composing the sub-tree representations corresponding to the given model, using its own compositional network. One can observe that the proposed model outperforms the best model by 1.1 F1-score on average. Moreover, it yields state-of-the art performance for 6 among the 9 available languages.

Finally, Table 3 compares the performance of different systems for a more realistic parsing scenario where the gold POS and morphological tags are unknown. For these experiments, we use the same tags as in (Björkelund et al., 2014)[2] obtained using the freely available tool MarMoT (Mueller et al., 2013). We compare our results with the same model as for the the gold tags experiences. Additionnaly, we compare our results with two recent models reporting results for the SPMRL Shared Task 2014. We see that the proposed model yields state-of-the art performance for 4 out of 9 available languages.

5 Conclusion

In this paper, we proposed to extend the parser introduced in (Legrand and Collobert, 2015) by learning morphological embeddings. We take advantage of a recursive procedure to propagate morphological information through the tree during the parsing process. We showed that using the morphological embeddings boosts the F1-score and allows to outperform the current state-of-the-art model on the SPMRL Shared Task 2014 corpus. Moreover, our approach yields state-of-the art performance for a majority of languages.

[2]The tags used are available here: `http://cistern.cis.lmu.de/marmot/models/CURRENT/`

Acknoledgments

This work was supported by NEC Laboratories America. We would like to thank Dimitri Palaz for our fruitful discussions and Marc Ferras for proofreading this paper.

References

Anders Björkelund, Ozlem Cetinoglu, Agnieszka Falenska, Richárd Farkas, Thomas Mueller, Wolfgang Seeker, and Zsolt Szántó. 2014. Introducing the ims-wrocław-szeged-cis entry at the SPMRL 2014 shared task: Reranking and morpho-syntax meet unlabeled data. *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine N-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*.

Brooke Cowan and Michael Collins. 2005. Morphology and reranking for the statistical parsing of spanish. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Greg Durrett and Dan Klein. 2015. Neural CRF parsing. In *Proceedings of the Association for Computational Linguistics*.

Daniel Fernández-González and Andr F. T. Martins. 2015. Parsing as reduction. In *Annual Meeting of the Association for Computational Linguistics ACL*.

David Hall, Greg Durrett, and Dan Klein. 2014. Less grammar, more features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Fei Huang and Alexander Yates. 2009. Distributional representations for handling sparsity in supervised sequence-labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural*

Language Processing of the AFNLP: Volume 1 - Volume 1.

Joël Legrand and Ronan Collobert. 2015. Joint RNN-based greedy parsing and word composition. In *Proceedings of ICLR*.

T. Mueller, H. Schmid, and H. Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Slav Petrov. 2010. Products of random latent variable grammars. In *NAACL-HLT*.

David C. Plaut and Geoffrey E. Hinton. 1987. Learning sets of filters using back-propagation. *Computer Speech and Language*.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Wolińsk, Alina Wróblewska, and Éric Villemonte De La Clergerie. 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the 4th Workshop on Statistical Parsing of Morphologically Rich Languages: Shared Task*.

Reut Tsarfaty, Djamé Seddah, Sandra Kbler, and Joakim Nivre. 2013. Parsing morphologically rich languages: Introduction to the special issue. *Computational Linguistics*.