Utilizing Word Embeddings for Part-of-Speech Tagging

Gábor Berend

Szegedi Tudományegyetem,
TTIK, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.,
e-mail:berendg@inf.u-szeged.hu

Abstract. In this paper, we illustrate the power of distributed word representations for the part-of-speech tagging of Hungarian texts. We trained CRF models for POS-tagging that made use of features derived from the sparse coding of the word embeddings of Hungarian words as signals. We show that relying on such a representation, it is possible to avoid the creation of language specific features for achieving reliable performance. We evaluated our models on all the subsections of the Szeged Treebank both using MSD and universal morphology tag sets. Furthermore, we also report results for inter-subcorpora experiments.

1 Introduction

Designing hand-crafted features for various natural language processing tasks, such as part-of-speech (POS) tagging or named entity recognition (NER) has a long going history [5,13]. Systems that build upon such (highly) language/task-specific features can often perform accurately, however, at the cost of losing their ability to work well across different languages and tasks. A further drawback of such approaches is that the human-powered design of features can be a time consuming and expensive task without any guarantees that the features work well under multiple circumstances or at all.

There is now a recent line of research gaining increasing popularity, which aims at building more general models that require no feature engineering at all but relying on large collections of (unlabeled) texts alone [2,3,4,9]. For the above reason these models can be regarded language independent, making them more likely to be applicable across languages.

Sparse coding aims at expressing observations as a sparse linear combination of 'basis vectors'[1] [7]. The goal of our work is to combine two popular approaches, i.e. sparse coding and distributed word representations.

In our work we propose a POS tagging architecture which was evaluated on the Szeged Treebank using MSD and universal morphology tag sets. We report

[1] The term basis vectors is used intuitively throughout the paper, as they need not be linearly independent

our POS tagging results on the levels of the six subcorpora the Szeged Treebank comprises of. Also, we evaluated our trained models in a cross-genre setting.

2 Related work

The line of research introduced in this paper relies on distributed word representations [1] and dictionary learning for sparse coding [7], both area having a substantial literature. This section introduces the most important previous work along these topics.

2.1 Distributed word representations

Distributed word representations provided by approaches such as `word2vec` [9] and `GloVe` [12], enjoy great popularity these days as they have been shown to accurately model the semantics of words [10]. This property makes them available to perform successfully in semantic and syntactic word analogy tasks. There exist previous results claiming that distributed word representations are also useful in the word analogy task in Hungarian (and other lower-resourced Central European languages) [8]. There exist a variety of approaches on how continuous word embeddings can be determined, e.g. [1,2,3,9,12].

The Polyglot [1] neural net architecture is one such possible alternative to determine word embeddings. In their proposed model, word embeddings were trained on the passages of Wikipedia, while preprocessing of texts was kept at a minimal level by not performing lowercasing or lemmatization. Applying such a generic approach for preprocessing not favoring any specific language makes this neural network architecture applicable for a variety of languages without any serious modifications. Indeed, the authors also made their pre-trained word embeddings for over 130 languages publicly available[2] providing basis for cross-, and multi-lingual experimentation. Since we wanted to give an approach that is not sensitive to the hyperparameters of the word embedding model, we applied those Polyglot word embedding vectors trained for Hungarian that are available for download at the Polyglot project website.

2.2 Sparse coding

Sparse coding has it roots in the computer vision community, and its usage is perhaps no so common in natural language processing literature. The general purpose of sparse coding is to express signals in the form of a *sparse* linear combinations of basis vectors, while the task of finding an appropriate set of basis vectors is referred to as *dictionary learning* problem [7]. Generally, given a data matrix $X \in \mathbb{R}^{k \times n}$ with its i^{th} column $\mathbf{x_i}$ representing the i^{th} k-dimensional signal, the task is to find $D \in \mathbb{R}^{k \times m}$ and $\alpha \in \mathbb{R}^{m \times n}$, such that the product of

[2] https://sites.google.com/site/rmyeid/projects/polyglot

matrices D and α approximates X. Mairal et al. [7] formalized this problem as an ℓ_1-regularized linear least-squares minimization of the form

$$\min_{D \in \mathcal{C}, \alpha} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \left(\|x_i - D\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1 \right),$$

with $\mathcal{C}$ being the convex set of matrices that comprise of column vectors having an ℓ_2 norm at most one, matrix D acts as the shared dictionary across the signals, and the columns of the sparse matrix α contains the coefficients for the linear combinations of each of the n observed signals. [7] describes an efficient algorithm for solving the above optimization that we also applied in our experiments[3].

3 Sequence labeling framework

This section introduces the sequence labeling framework we employed for POS tagging. During our experiments the main source of features for the tokens in a sentence was the dictionary learning based sparse coding of their word embedding vector. Once the dictionary matrix D is given α_i, the sparse linear combination coefficients for a word embedding vector w_i, can be determined efficiently by solving the kind of minimization problem described in Section 2.2. The way we turned these sparse coefficients into features was that we regarded those indices of α_i as features that had a non-zero value, i.e. $f(w_i) = \{j : \alpha_i[j] \neq 0\}, \alpha_i[j]$ denoting j^{th} coefficient stored in the sparse vector α_i. It can be illustrative if we check out the kind of features that got determined for semantically related words. Table 1 includes such a set of words and their corresponding features. In Table 1 any feature ID appearing more than got boldfaced.

The only language dependent feature we made use of was the identity of words. For the calculation of this feature we performed no preprocessing, i.e. the words were not lemmatized and even their capitalization was left unchanged.

Word	Sparse features induced
kéz (hand)	{144, 218, **309**, 472, 713, 870, **916**}
láb (leg)	{138, 186, **250**, **309**, 324, 583, 626, 796, **948**}
fej (head)	{101, **250**, 271, **309**, 516, **783**, **916**, **948**}
törzs (trunk)	{81, **309**, **783**, 867, 948}
csukló (wrist)	{84, 194, **309**, 607, 815, 957}

Table 1: Example words all being body parts and the sparse features induced for them. Features with multiple occurrences across words are in **bold** typeface. Within parenthesis are the English equivalent of the Hungarian example words.

When assigning features to a target word at some position within a sentence, we determined the same set of feature functions for the target word itself and its

[3] `http://spams-devel.gforge.inria.fr/`

neighboring words of window size 1. We then used the previously described set of features in a linear chain CRF [6], using the CRFsuite implementation [11]. The coefficients for ℓ_1 and ℓ_2 regularization were set to 1.0 and 0.001, respectively.

4 Results and discussion

We evaluated our proposed POS tagging framework on the Szeged Treebank [14] which has six subcorpora, namely text related to *computer*s, *law*, *literature*, short news (referenced as *newsml*), *newspaper* articles and *student* writing. The performance of our POS tagger models are expressed as the fraction of correctly tagged tokens (per-token) evaluation and as a fraction of the correctly tagged sentences (per-sentence) evaluation when a sentence is regarded as correct if all the tokens it comprises are tagged correctly. Evaluation was performed according to the reduced tag set of the MSD v2.5 and the universal morphologies as well. In the two distinct tag sets, we faced a 93-class and a 17-class sequence classification problem, respectively. The dictionary learning approach we made use of relied on two parameters, the dimensionality of the basis vectors and the regularization parameter effecting the sparsity of the coefficients in α. We chose the former parameter to be 1024 and the latter to be 0.4, nevertheless we should also add the general tendencies remained the same when we chose other pairs of parameters.

The first factor that could influence the performance of our approach is the coverage of the word embedding vectors employed, i.e. what extent of the training/test tokens/word forms do we have a distributed representation determined for. Table 2 includes these information. We can see that due to the morphological richness of Hungarian, the word form coverage of the roughly 150,000 word embedding vectors we had access to is relatively low (around 60%) for all the domains in the treebank. Due to the Zipfian distribution of word frequencies, however, we could experience a much higher (almost 90%) coverage for all the domains in the treebank on the level of tokens. It is interesting to see that student writings has one of the lowest word form coverage, while it is among the genres with the highest token coverage. It might indicate that student writing is not as elaborate and standardized as news writing for instance.

	Training		Test		Average
Domain	Tokens	Word forms	Tokens	Word forms	Tokens
computer	88.54% (4)	60.13% (3)	88.76% (4)	69.42% (3)	88.59% (4)
law	86.04% (6)	58.80% (4)	86.10% (6)	65.15% (5)	86.06% (6)
literature	90.12% (1)	58.56% (5)	89.97% (1)	68.58% (4)	90.09% (1)
newsml	87.67% (5)	63.15% (2)	87.72% (5)	69.85% (2)	87.68% (5)
newspaper	89.22% (3)	63.69% (1)	89.25% (3)	72.48% (1)	89.22% (3)
student	89.68% (2)	54.32% (6)	89.70% (2)	63.04% (6)	89.69% (2)
Total	**88.59%**	—	**88.61%**	—	**88.60%**

Table 2: The token and word form coverages of the Polyglot word embeddings on the Szeged Treebank. In parenthesis are the ranks for a given domain.

Regarding our POS tagging results, in all our subsequent tables, we report three numbers per each cross-domain evaluation. The three numbers refer to the three kinds of experiments below:

1. only word identity features are utilized,
2. both word identity and sparse coding-derived features are utilized,
3. only sparse coding-derived features are utilized.

Next, we present our evaluation across the six distinct categories of Szeged Treebank according to the reduced MSD v2.5 tag set consisting of 93 labels. Table 3 and Table 4 contains our results depending on whether accuracies were calculated on the per-token or per-sentence level, respectively.

Train \ Test	computer	law	literature	newsml	newspaper	student
computer	88.47%	80.00%	74.11%	81.37%	79.70%	76.55%
	92.57%	88.19%	83.86%	88.75%	89.28%	82.84%
	90.07%	85.91%	80.73%	86.66%	86.49%	80.34%
law	76.35%	93.52%	64.89%	70.61%	72.87%	67.70%
	86.24%	95.47%	75.65%	83.32%	85.41%	76.83%
	83.95%	92.69%	73.06%	80.90%	82.84%	74.48%
literature	73.63%	68.01%	88.17%	64.16%	75.21%	84.71%
	85.81%	82.51%	91.65%	81.40%	86.97%	88.66%
	83.34%	80.79%	89.15%	79.03%	84.65%	85.81%
newsml	77.91%	76.64%	67.57%	93.28%	77.94%	70.88%
	86.73%	86.02%	76.72%	95.79%	87.20%	77.73%
	84.57%	84.37%	75.27%	93.79%	85.11%	75.43%
newspaper	82.21%	80.90%	79.68%	86.61%	85.78%	81.00%
	89.26%	88.75%	86.48%	91.48%	91.32%	85.69%
	87.04%	86.44%	84.02%	88.77%	88.94%	82.70%
student	75.27%	70.65%	82.74%	72.71%	77.80%	91.53%
	85.15%	82.50%	88.18%	83.45%	87.23%	93.21%
	82.24%	79.32%	85.42%	80.12%	84.11%	89.80%

Table 3: Per-token cross-evaluation accuracies across the subcorpora of Szeged Treebank using a reduced tag set of MSD version 2.5 consisting of 93 labels.

Subsequently, we evaluated our models according to all the possible combinations of the subcorpora relying on the coarser-level universal morphologies tag set which includes 17 POS tags. Results for the per-token and sentence-level evaluations are present in Table 5 and Table 6, respectively.

Comparing the results when evaluating according to the MSD tagset and the universal morphologies, we can observe that better results were achieved when evaluation took place according to the universal morphologies. This is not so surprising, however, as the task was simpler in the latter case, i.e. we faced a 17-class sequence classification problem, opposed to the 93-class problem for the MSD case.

Train \ Test	computer	law	literature	newml	newspaper	student
computer	21.21%	3.79%	8.31%	2.92%	6.16%	6.39%
	30.93%	12.71%	18.88%	11.35%	18.20%	12.79%
	21.26%	9.54%	13.87%	8.42%	12.32%	9.54%
law	4.64%	31.17%	3.28%	0.81%	3.22%	3.01%
	13.37%	41.08%	6.68%	4.74%	10.90%	7.25%
	9.57%	24.38%	5.25%	3.68%	7.44%	5.50%
literature	3.70%	1.50%	36.43%	0.40%	6.26%	19.76%
	11.00%	5.08%	43.86%	2.62%	14.60%	26.49%
	8.24%	3.79%	34.91%	2.12%	10.09%	18.64%
newsml	4.64%	2.23%	3.22%	42.56%	4.79%	3.35%
	13.37%	8.97%	7.27%	50.68%	12.42%	7.23%
	9.92%	6.85%	6.68%	35.30%	8.58%	6.01%
newspaper	8.68%	4.62%	14.38%	6.61%	12.27%	11.75%
	19.14%	12.24%	25.03%	14.52%	23.36%	17.59%
	12.97%	9.08%	19.76%	10.14%	16.97%	13.07%
student	3.55%	0.99%	22.08%	0.76%	6.21%	40.09%
	10.71%	5.50%	31.58%	5.14%	14.41%	45.79%
	7.70%	3.37%	24.05%	3.23%	9.43%	31.49%

Table 4: Per-sentence cross-evaluation accuracies across the subcorpora of Szeged Treebank using a reduced tag set of MSD version 2.5 consisting of 93 labels.

Train \ Test	computer	law	literature	newsml	newspaper	student
computer	90.66%	84.05%	78.54%	83.62%	81.84%	83.28%
	94.56%	91.63%	88.38%	91.63%	91.59%	90.52%
	92.35%	89.32%	86.29%	90.21%	89.30%	88.35%
law	78.18%	96.07%	70.07%	72.91%	75.94%	73.81%
	88.18%	97.67%	82.38%	86.90%	87.00%	84.38%
	86.43%	95.65%	80.35%	85.76%	85.51%	82.21%
literature	76.70%	75.64%	91.54%	66.17%	78.19%	88.90%
	87.54%	87.87%	95.16%	82.38%	90.05%	93.36%
	85.70%	85.69%	92.92%	80.49%	88.11%	91.23%
newsml	79.83%	81.36%	69.71%	94.50%	79.62%	75.02%
	89.51%	90.42%	85.19%	97.07%	90.70%	85.62%
	87.88%	88.96%	83.30%	95.58%	88.53%	83.33%
newspaper	84.08%	85.89%	83.48%	88.29%	88.38%	86.51%
	91.43%	91.93%	91.23%	93.59%	94.01%	91.96%
	89.89%	90.28%	89.55%	91.32%	91.85%	89.61%
student	77.49%	75.77%	85.41%	69.89%	79.61%	93.88%
	88.73%	87.97%	92.08%	85.74%	90.56%	96.04%
	85.83%	84.45%	90.28%	82.69%	88.22%	94.04%

Table 5: Per-token cross-evaluation accuracies across the subcorpora of Szeged Treebank using the universal morphology tag set.

Applying either kind of evaluation, the domain of newspapers seems to be the hardest one in the intra-domain evaluation, as the lowest accuracies are reported here. Also, we can notice that the *literature* and *student* domains are the most different from the others, as training on these corpora and evaluating against some other yields the biggest performance drops. Although *literature* and *student* writing being substantially different from all the other genres, they seem to be similar to each other, as the performance gap when training on one of these domains and evaluating on the other has milder performance gaps compared to other scenarios.

It can be clearly seen that models using features for both the word identities and sparse coding has the best results often by a large margin. It is not surprising as this model had access to the most information. When comparing the results of the models which either solely relied on word identity or sparse coding features, it is interesting to note that the model not relying on the identity of words ar all, but the sparse coding features alone, tends to perform better. A final important observation to take is that when sparse coding features are employed, domain differences seem to be expressed less, i.e. the performance drops in cross-domain evaluation settings tend to lessen.

Train \ Test	computer	law	literature	newml	newspaper	student
computer	26.64% 41.54% 29.26%	8.25% 23.91% 17.12%	13.85% 28.63% 22.88%	5.24% 20.42% 13.77%	10.66% 26.49% 19.62%	16.69% 31.89% 24.64%
law	5.97% 18.55% 13.37%	47.93% 63.28% 42.48%	5.49% 14.35% 11.96%	1.31% 7.87% 5.95%	4.50% 14.69% 12.23%	6.13% 15.59% 12.11%
literature	5.33% 17.56% 12.93%	3.53% 14.11% 9.75%	48.34% 60.51% 48.87%	0.61% 5.40% 3.53%	9.10% 22.70% 17.58%	31.23% 45.29% 35.07%
newml	6.36% 19.39% 13.32%	5.29% 17.84% 13.74%	5.17% 19.63% 16.14%	48.41% 59.51% 44.13%	7.58% 20.76% 14.88%	6.79% 17.73% 14.04%
newspaper	10.71% 27.97% 19.68%	8.82% 23.55% 17.43%	21.44% 39.52% 32.89%	12.15% 25.67% 17.35%	19.95% 36.35% 27.01%	23.16% 36.92% 27.84%
student	6.22% 17.07% 12.33%	2.75% 14.06% 9.60%	29.03% 44.82% 36.99%	1.01% 7.56% 4.74%	9.67% 24.08% 18.63%	50.89% 62.46% 48.76%

Table 6: Per-sentence cross-evaluation accuracies across the subcorpora of Szeged Treebank using the universal morphology tag set.

5 Conclusion

In this paper, we described our CRF-based POS-tagging model relying on the sparse coding of distributed word representations. We evaluated our proposed method on the subsections of the Szeged Treebank and found that the sparse coding derived features help to lessen the domain differences in cross-genre evaluation settings. We also found that relying on sparse coding features alone, it is possible to obtain better tagging accuracies than using word identity features and that combining the two sources of information can yield the best accuracies.

References

1. Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.
2. Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March 2003.
3. Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, New York, NY, USA, 2008. ACM.
4. Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November 2011.
5. Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. Named entity recognition through classifier combination. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 168–171, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.
6. John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.
7. Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, 11:19–60, March 2010.
8. Márton Makrai. Comparison of distributed language models on medium-resourced languages. *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*, pages 22–33, 2015.
9. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.
10. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.
11. Naoaki Okazaki. CRFsuite: a fast implementation of Conditional Random Fields (CRFs), 2007.

12. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, 2014.
13. Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.
14. Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. Hungarian dependency treebank. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).