L'optimisation des plongements de mots pour le français : une application de la classification des phrases

Jungyeul Park
CONJECTO, 74 rue de Paris, 35000 Rennes, France
http://www.conjecto.com

RÉSUMÉ

Nous proposons trois nouvelles méthodes pour construire et optimiser des plongements de mots pour le français. Nous utilisons les résultats de l'étiquetage morpho-syntaxique, de la détection des expressions multi-mots et de la lemmatisation pour un espace vectoriel continu. Pour l'évaluation, nous utilisons ces vecteurs sur une tâche de classification de phrases et les comparons avec le vecteur du système de base. Nous explorons également l'approche d'adaptation de domaine pour construire des vecteurs. Malgré un petit nombre de vocabulaires et la petite taille du corpus d'apprentissage, les vecteurs spécialisés par domaine obtiennent de meilleures performances que les vecteurs hors domaine.

ABSTRACT

Optimization of Word Embeddings for French : an Application of Sentence Classification.

We propose three novel methods for building word embeddings for French. We use results from part of speech tagging, detection of multiword expressions and lemmatization for a continuous vector space. For evaluation, we use these embedding vectors in a sentence classification task and compare them with the baseline embedding vector. We also explore domain adaptation approach for building embedding vectors, in which even with a small number of vocabularies and the small size of the training corpus, in-domain embeddings perform better than out-domain embeddings.

MOTS-CLÉS : Plongements de mots, catégorie grammaticale, expressions multi-mots, lemme, classification des phrases, français.

KEYWORDS: word embeddings, part of speech, multiword expression, lemma, sentence classification, French.

1 Introduction

Word embedding techniques have prevailed in natural language processing (NLP) and have obtained impressive results in several areas (Erhan *et al.*, 2010). Word embeddings are trained on word co-occurrence in text, and can capture semantic information about words and their meanings. Word2vec (Mikolov *et al.*, 2013) and GloVe (Pennington *et al.*, 2014) have been proposed to learn a distributed representation for words in a continuous vector space. Word embedding vectors with subwords were also presented to improve embedding quality (Luong *et al.*, 2013; Bojanowski *et al.*, 2017). There are still many *2vec-style variations. Multilingual embedding (Ammar *et al.*, 2016) or embedding with polysemy (Arora *et al.*, 2016) have also been presented as one extension of word embedding

techniques. At syntactic level, Levy & Goldberg (2014) proposed dependency-based word embeddings, and He *et al.* (2018) developed even further to encode the syntactic-aware context of entities. As proposed in previous work word embedding techniques have rapidly been used in a large range of applications in NLP over the last years.

In this paper, we present three methods for building and optimizing embedding vectors for French. We use POS tagging, multiword expression detection and lemmatization of words to optimize a word embedding task. We extrinsically evaluate the proposed method for embedding vectors by using a sentence classification task. The current work differs from previously proposed opinion mining (or sentiment analysis) (*inter alia*, Pang & Lee (2008)) where the classification task in previous work is mainly based on the document. For sentence classification, Socher *et al.* (2012) shows phrase fragments classification using a recursive neural network model. More recently, Dernoncourt *et al.* (2017a) proposes the sequential sentence classification task by adding a sentence label prediction layer.

The main contribution of this paper is as follows. First, we explore linguistically motivated methods to optimize word embeddings for French. Secondly, we propose a new data set for the sentence classification task for French. In addition to embedding optimization and sentence classification, we introduce a domain adaptation approach by selecting the training corpus for embeddings. We find several practical facts in word embeddings including the effects on the size of the training corpus, the size of vocabularies, and their relatedness with the task.

2 Optimizing Word Embeddings for French

This section describes three linguistically motivated methods for optimizing word embedding vectors for French : POS-aware, MWE-aware and lemma-aware embeddings. Instead of directly using surface forms of the word, we first use a pair of word and part of speech (POS) to disambiguate possible polysemy. Secondly, since words in multiword expressions (MWEs) can give more senses when they are bonded together, MWEs are dealt as a single unit. Thirdly, we use lemma forms in order to avoid sparsity because of rarely used inflected words. While the current methods have been proposed and used in the bag-of-words model, to our best knowledge there are no previous efforts for word embeddings to optimize the vector space model for French. We note that, while Ferré (2017) proposed complex terms for embeddings, he worked on embeddings for English.

Previously, character (Chrupała, 2013; Chen *et al.*, 2015; Wieting *et al.*, 2016) syllable (Yu *et al.*, 2017; Choi *et al.*, 2017), and subword (Mikolov *et al.*, 2012; Bojanowski *et al.*, 2017; Pinter *et al.*, 2017; González-Gallardo & Torres-Moreno, 2017; Stratos, 2017) embedding techniques have been proposed to enrich word vectors.

2.1 POS-aware embedding

POS tagging is one of the simplest, but the most important and well-studied tasks in NLP. Various supervised and unsupervised approaches have been proposed for POS tagging. Besides traditional rule-based approaches, in which POS dictionaries and manually crafted rules (*e.g.* syntagmatic patterns) are required, there are several supervised learning methods that can learn from POS tagged data such as transformation-based learning (Brill, 1995), hidden Markov models (HMM) (Kupiec,

... l'/O *arbitraire*/O *de*/O *la*/O *démesure*/O *veut*/O *susciter*/O *à*/B-MWE *tout*/I-MWE *prix*/I-MWE *.*/O

FIGURE 1 – MWE-annotated corpus from the French treebank

1992), maximum entropy models (Ratnaparkhi, 1996), Conditional random fields (CRF) (Lafferty *et al.*, 2001), etc. POS labels can give an additional information source for the word, especially some cases for homonyms and homographs. For example, while a noun *joue* ('cheek') and a verb *joue* ('play') would be considered as the same word in the general word embedding, *joue*/N and *joue*/V can be distinguishable. Therefore, we use a pair of the word and its POS label as a single unit for embeddings. Training a model with POS labels has also been proposed in statistical machine translation as one of factored training.

Any sequence labeling algorithm easily achieves state of arts results for POS tagging. We train and evaluate POS tagging using TnT (Brants, 2000) for an HMM and Wapiti (Lavergne *et al.*, 2010) for CRFs. The French treebank (Abeillé *et al.*, 2003) is used for training and evaluation of POS tagging based on a corpus split proposed in Seddah *et al.* (2013). We obtain 96.49% and 97.76% for the HMM and CRFs, respectively. Therefore, we use CRFs POS tagging results to preprocess the corpus for building POS-aware embedding, and accordingly the classification data set.

2.2 MWE-aware embedding

While multiword expressions (MWE) have been considered as a pain in the neck for natural language processing (Sag *et al.*, 2002), it also shows their importance in NLP. For French MWEs, Daille (2003) studied in the context of terminology, and Daille *et al.* (2004) and Morin & Daille (2010) presented MWEs in English-French translation alignment from the comparable corpus. Green *et al.* (2011) used tree substitution grammars to improve MWE identification and constituent parsing results. Dubremetz & Nivre (2014) created MWE data using the French Europarl corpus and corpus together with the lexicon of local grammars (Gross, 1975) for MWE detection by classification. Words in MWEs can give more senses when they are co-occurred instead of being processed separately. Therefore, we detect MWEs and deal with them as a single unit. For example, *au* from *au contraire* and *au cours de* are different in MWE-aware embedding because there are *au-contraire* and *au-cours-de* are also listed in addition to a single word *au*. Their distributions are calculated independently without considering *au* and *contraire*, or *au cours* and *de*.

The French treebank contains MWE annotation, in which phrase labels share the same label names with POS labels, for example, [**P** [P *D'*][P *après*]]. For MWE-aware embedding, we convert the treebank sentences into the sequence labeled sentences using the BIO format such as `B-MWE` and `I-MWE` as described in Figure 1. B-I-O stands for beginning-inside-outside of MWEs. We train bi-directional long-short-term-memory recurrent neural networks (bi-LSTM RNN) (Graves & Schmidhuber, 2005) using NeuroNER (Dernoncourt *et al.*, 2017b), and obtain up to 79.75% F1 score (79.47% in average for 5 runs). The French treebank is still used for training and evaluation of MWE detection. We note that we also train and evaluate MWE data using CRFs with ± 2 word and POS context information as a feature set, in which we obtain only 74.61% F1 score. Therefore, we use bi-LSTM RNN MWE detection to preprocess the corpus, and put together words of MWEs as a single unit for building MWE-aware embedding.

Wikipedia	Europarl	New crawls	Common crawl	Giga	in-domain	total
675M	64M	223M	89M	793M	664M	2.5B

TABLE 1 – The size (# of token) of corpora for embedding

2.3 Lemma-aware embedding

French is a moderately inflected language. [1] For example, a verb *avoir* ('have') can be inflected depending on person, number, mood, and tense : *ai*, *eus*, *avais*, *aurai*, *aie*, *eusse*, *aurais* for the first-person singular. Previously, Flemm used a rule-based method for lemmatization with POS-tagged results (Namer, 2000). L*efff* is a semi-automatically developed morphological and syntactic lexicon for French (Sagot, 2010) and there are systems based on L*efff* for lemmatization. We use a canonical form of the word (lemma) as in its basic unit of word embeddings, and we refer it as lemma-aware embedding. In lemma-aware embedding, different inflected forms such as *ai*, *eus*, *avais* are equally dealt with as *avoir*.

For lemma-aware embedding, we use a pipeline of TreeTagger (Schmid, 1994) and Flemm (Namer, 2000) for lemmatization. For unknown lemmas, especially for proper nouns, we directly use the surface form. Since there is currently no available evaluation data for lemmatization, we do not evaluate lemmatization results. The original French treebank contains such lemmatization information in the XML format. For example, a word *a* is annotated as a verb along with morphosyntactic properties including its lemma : `<w cat="V" ee="V-P3s" ei="VP3s" lemma="avoir" mph="P3s" subcat="">a</w>`. We leave the evaluation of lemmatization for future work.

2.4 Building word embeddings

The corpora for embedding include as follows :

1. Wikipedia, Europarl, and News Crawls for monolingual French,
2. French-side from the parallel corpus such as Common Crawl and Giga French-English, and
3. our in-domain corpus (described in § 3.1).

The size of these corpora (number of tokens) is summarized in Table 1. We use symbol normalization and tokenization schemes for French in Moses (Koehn *et al.*, 2007). Then, we post-edit tokenization errors such as a misuse of the apostrophe character. Since there are several tokenization error cases especially for a contraction, we build heuristic regular expressions to correct them. For sentence boundaries, we use TreeTagger's sentence boundary detection. All characters are lowercased for embeddings. Figure 2 presents sentence examples of the initial baseline, MWE-aware, and lemma-aware corpus for different embeddings. We build 300 dimension skip-gram embeddings with default options : 0.05 learning rate, and 5 for the size of the context window, etc.

Technically, we do not deal with proposed optimization methodologies putting together because they are contradictory. While POS-aware and MWE-aware embedding techniques increase sparsity by giving more number of vocabulary in the vector, lemma-aware embedding vector is intended to avoid sparsity.

1. `https://en.wikipedia.org/wiki/French_language`, accessed on January 12, 2018.

Initial : ... *l'* ␣ *arbitraire* ␣ *de* ␣ *la* ␣ *démesure* ␣ *veut* ␣ *susciter* ␣ *à* ␣ *tout* ␣ *prix* ␣ .
POS-aware : ... *l'/D* ␣ *arbitraire/N* ␣ *de/P* ␣ *la/D* ␣ *démesure/N* ␣ *veut/V* ␣ *susciter/V* ␣ *à/P* ␣ *tout/D* ␣ *prix/N* ␣ *./PONCT*
MWE-aware : ... *l'* ␣ *arbitraire* ␣ *de* ␣ *la* ␣ *démesure* ␣ *veut* ␣ *susciter* ␣ *à-tout-prix* ␣ .
Lemma-aware : ... *le* ␣ *arbitraire* ␣ *de* ␣ *le* ␣ *démesure* ␣ *vouloir* ␣ *susciter* ␣ *à* ␣ *tout* ␣ *prix* ␣ .

FIGURE 2 – Example of the embedding corpus

3 Sentence Classification

We extrinsically evaluate the proposed optimization methods for word embedding by using the sentence classification task. First, we build sentence classification data, and then present classification results using various embedding settings and corpus sizes. We also explore the domain adaptation approach by using the in-domain corpus to build embedding vectors.

3.1 Sentence classification data and in-domain corpus

We build the sentence classification data set for French using opportunities : by opportunities, we deal with financial contract opportunities appeared in the municipal debriefing report. We download municipal debriefing reports from the city council all over France.

Our in-domain corpus consists of 886K documents and 664M tokens from these municipal debriefing reports. For classification, about 4,000 sentences are manually annotated either positive or negative to build the classification model. The classification data set is described in detail in Park *et al.* (2018). As described, all data sets are POS-annotated for POS-aware embedding, MWE-detected for MWE-aware embedding, and lemmatized for lemma-aware embedding experiments.

3.2 Classification experiments

For evaluation, we use several different embedding settings : baseline (based on inflected forms without POS labels, MWE detection or lemma), POS-aware (P-embedding), MWE-aware (M-embedding) and lemma-aware (L-embedding). We also build the embedding vector using the different size of the corpus : IN using the in-domain *relatively* small corpus and OUT using the large out-domain corpus from various sources for a domain adaptation approach. We use fastText (Joulin *et al.*, 2016) for building word embedding and classification. For embeddings, we build 300 dimension skip-gram models with default options. For classification, we use 1.0 for the learning rate, 25 epochs, and proposed pre-trained word vectors.

Table 2 shows classification results (accuracy) based on the different configurations alongside the size of vocabularies in embeddings. For a comparison purpose, we perform sentence classification using a pre-trained embedding vector for French provided by Bojanowski *et al.* (2017) (WIKI). It is a 300 dimension skip-gram model as ours. We also report a classification result "without" word embeddings (NONE) to show the effects of embedding vectors in classification.

Based on experiment results, we find following several practical facts. Using word embeddings improves classification results for all cases. Classification results are improved as the size of voca-

	IN	OUT	WIKI	NONE
baseline embedding	0.918 (0.4M)	0.909 (1.3M)	0.906 (1.1M)	0.901
P-embedding	0.919 (0.5M)	0.914 (1.8M)	-	0.912
M-embedding	0.920 (0.6M)	0.911 (1.7M)	-	0.909
L-embedding	0.901 (0.3M)	0.896 (1.1M)	-	0.891

TABLE 2 – Results based on the different embedding setting and the corpus size. We also provide the size of vocabularies. NONE is for the classification result without word embeddings. All data sets are POS-annotated, MWE-detected, and lemmatized for each experiment.

	IN	OUT	WIKI	NONE
baseline embedding	0.937	0.934	0.934	0.929
P-embedding	0.937	0.935	-	0.930
M-embedding	0.942	0.937	-	0.934
L-embedding	0.932	0.931	-	0.927

TABLE 3 – Results based on the bigram feature. All data sets are POS-annotated, MWE-detected, and lemmatized for each experiment as before.

bularies increase. Especially, "in-domain" information plays an important role regardless of its size in embeddings. Even with a small number of vocabularies and the small size of the training corpus, in-domain embeddings always outperform out-domain embeddings. [2]

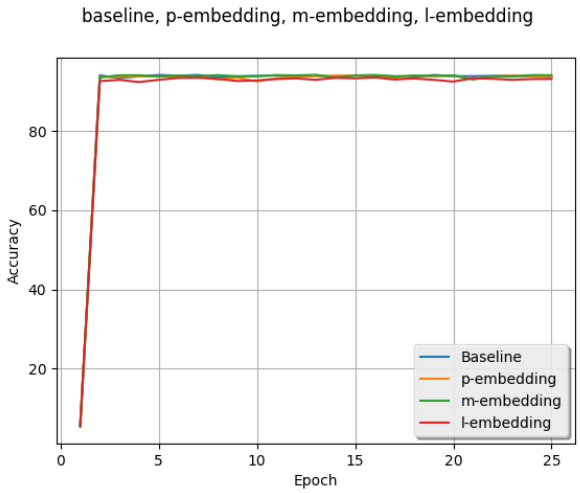
Results also show that context information matters to improve overall results such as in MWE-embeddings. This finding indicates that considering context as in MWE-aware embedding affect classification results. Therefore, co-occurred words should be processed together.

Finally, since context information matters, we extend experiments using a simple bigram feature. Wang & Manning (2012) already used bigram features to improve classification results for naive Bayes and support vector machines. Table 3 shows classification results with the bigram feature. Figure 3 shows results for each epoch using the bigram feature. While results for all embedding are converged in a very early stage, M-embedding can yield better results for almost all epochs (Figure 3a). We also compare classification results using embedding by in- and out-domain copora for m-embeddings (Figure 3b) where embeddings with the in-domain corpus outperform the out-domain corpus.

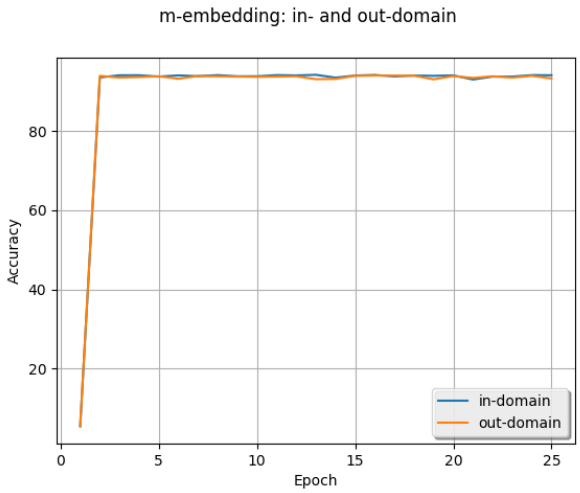
4 Discussion and Conclusion

We proposed several methods for optimizing word embeddings for French by using POS tagging, MWE detection and lemmatization. We used these embedding vectors in the sentence classification task, and MWE-aware in which the size of vocabularies are relatively large to other embeddings improved the classification result. To extend current MWE-aware embedding, we can consider formulaic sequences and named entities (Brooke *et al.*, 2017) in addition to MWEs. We leave these extensions for future work. While reported results show a minor improvement, they confirm our

2. We note that Fabre *et al.* (2014) also used the small size of the specialized corpus, which is similar to our in-domain corpus embeddings.



(a) Comparison of different embeddings



(b) Comparison of in-/out-domain corpora

FIGURE 3 – Results of each epoch : (a) for POS-aware (P-embedding), MWE-aware (M-embedding) and lemma-aware (L-embedding) using the in-domain corpus and the bigram feature, and (b) for MWE-aware (M-embedding) using the in/out-domain corpora and the bigram feature

intuition that incorporated context information in embeddings either linguistically motivated (MWEs) or not (bigram features) is important. As described, "in-domain" information played an important role, which should be well adapted to the proposed task. In-domain embeddings using a small number of vocabularies and the small size of the training corpus (roughly a third of the out-domain corpus) outperformed out-domain embeddings for all cases. The word embedding vectors, preprocessed source text files including the preprocessing script, and the sentence classification data for French are publicly available at `https://github.com/jungyeul/taln2018`.

Remerciements

We would like to thank the anonymous reviewers for their suggestions and comments.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a Treebank for French. In *Treebanks*, p. 165–188. Kluwer.

AMMAR W., MULCAIRE G., TSVETKOV Y., LAMPLE G., DYER C. & SMITH N. A. (2016). Massively Multilingual Word Embeddings. *http ://arxiv.org/abs/1602.01925*.

ARORA S., LI Y., LIANG Y., MA T. & RISTESKI A. (2016). Linear Algebraic Structure of Word Senses, with Applications to Polysemy. *http ://arxiv.org/abs/1601.03764*.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, **5**, 135–146.

BRANTS T. (2000). TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, p. 224–231, Seattle, Washington, USA : Association for Computational Linguistics.

BRILL E. (1995). Transformation-Based-Error-Driven Learning and Natural Language Processing : A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, **21**(4), 543–566.

BROOKE J., SNAJDER J. & BALDWIN T. (2017). Unsupervised Acquisition of Comprehensive Multiword Lexicons using Competition in an n-gram Lattice. *Transactions of the Association for Computational Linguistics*, **5**, 455–470.

CHEN X., XU L., LIU Z., SUN M. & LUAN H. (2015). Joint learning of character and word embeddings. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, p. 1236–1242, Buenos Aires, Argentina : AAAI Press.

CHOI S., KIM T., SEOL J. & LEE S.-G. (2017). A Syllable-based Technique for Word Embeddings of Korean Words. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, p. 36–40, Copenhagen, Denmark : Association for Computational Linguistics.

CHRUPAŁA G. (2013). Text segmentation with character-level text embeddings. In *Proceedings of ICML 2013 workshop on Deep Learning for Audio, Speech and Language Processing*, Atlanta, GA.

DAILLE B. (2003). Conceptual Structuring through Term Variations. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, p. 9–16, Sapporo, Japan : Association for Computational Linguistics.

DAILLE B., DUFOUR-KOWALSKI S. & MORIN E. (2004). French-English Multi-word Term Alignment Based on Lexical Context Analysis . In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, p. 919–922, Lisbon, Portugal : European Language Resources Association (ELRA).

DERNONCOURT F., LEE J. Y. & SZOLOVITS P. (2017a). Neural Networks for Joint Sentence Classification in Medical Paper Abstracts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, p. 694–700, Valencia, Spain : Association for Computational Linguistics.

DERNONCOURT F., LEE J. Y. & SZOLOVITS P. (2017b). NeuroNER : an easy-to-use program for named-entity recognition based on neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 97–102, Copenhagen, Denmark : Association for Computational Linguistics.

DUBREMETZ M. & NIVRE J. (2014). Extraction of Nominal Multiword Expressions in French. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, p. 72–76, Gothenburg, Sweden : Association for Computational Linguistics.

ERHAN D., BENGIO Y., COURVILLE A., MANZAGOL P.-A., VINCENT P. & BENGIO S. (2010). Why Does Unsupervised Pre-training Help Deep Learning ? *Journal of Machine Learning Research*, **11**, 625–660.

FABRE C., HATHOUT N., SAJOUS F. & TANGUY L. (2014). Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille. In *Actes de la 2ème édition de l'atelier SemDis (SemDis 2014)*, p. 266–279, Marseille, France.

FERRÉ A. (2017). Représentation de termes complexes dans un espace vectoriel relié à une ontologie pour une tâche de catégorisation. In *Actes de Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA 2017)*, Caen, France.

GONZÁLEZ-GALLARDO C.-E. & TORRES-MORENO J.-M. (2017). Sentence Boundary Detection for French with Subword-Level Information Vectors and Convolutional Neural Networks. In *Proceedings of the International Conference on Natural Language, Signal and Speech Processing (ICNLSSP 2017)*, p. 80–84, Casablanca, Morocco.

GRAVES A. & SCHMIDHUBER J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, **18**(5-6), 602–610.

GREEN S., DE MARNEFFE M.-C., BAUER J. & MANNING C. D. (2011). Multiword Expression Identification with Tree Substitution Grammars : A Parsing tour de force with French. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, p. 725–735, Edinburgh, Scotland, UK. : Association for Computational Linguistics.

GROSS M. (1975). *Méthodes en syntaxe*. Hermann.

HE Z., CHEN W., LI Z., ZHANG M., ZHANG W. & ZHANG M. (2018). SEE : Syntax-aware Entity Embedding for Neural Relation Extraction. In *Proceedings of The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, New Orleans, Louisiana.

JOULIN A., GRAVE E., BOJANOWSKI P., DOUZE M., JÉGOU H. & MIKOLOV T. (2016). FastText.zip : Compressing text classification models. *http ://arxiv.org/abs/1612.03651*.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (2007). Moses : Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, p. 177–180, Prague, Czech Republic : Association for Computational Linguistics.

KUPIEC J. (1992). Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language*, **6**(3), 225–242.

LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, p. 282–289, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical Very Large Scale CRFs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 504–513, Uppsala, Sweden : Association for Computational Linguistics.

LEVY O. & GOLDBERG Y. (2014). Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 302–308, Baltimore, Maryland : Association for Computational Linguistics.

LUONG T., SOCHER R. & MANNING C. D. (2013). Better Word Representations with Recursive Neural Networks for Morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, p. 104–113, Sofia, Bulgaria : Association for Computational Linguistics.

MIKOLOV T., SUTSKEVER I., ANOOP D., HAI-SON L., STEFAN K. & JAN C. (2012). Subword Language Modeling with Neural Networks.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. BURGES, L. BOTTOU, M. WELLING, Z. GHAHRAMANI & K. Q. WEINBERGER, Eds., *Advances in Neural Information Processing Systems 26*, p. 3111–3119. Curran Associates, Inc.

MORIN E. & DAILLE B. (2010). Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*, **44**(1-2), 79–95.

NAMER F. (2000). Un analyseur Flexionnel de Français à base de règles. *Revue TAL*, **41**(2), 523–548.

PANG B. & LEE L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, **2**(1-2), 1–135.

PARK J., DELAHAYE E. & BOVYN R. (2018). Detectio : Web Interface for Building Sentence Classification and User Recommendation Data. In *Proceedings of the 15th ESWC 2018 (INDUSTRY TRACK)*, Crete, Greece.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). GloVe : Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543, Doha, Qatar : Association for Computational Linguistics.

PINTER Y., GUTHRIE R. & EISENSTEIN J. (2017). Mimicking Word Embeddings using Subword RNNs. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 102–112, Copenhagen, Denmark : Association for Computational Linguistics.

RATNAPARKHI A. (1996). A Maximum Entropy Model for Part-Of-Speech Tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 133–142, Philadelphia, Pennsylvania, USA.

SAG I. A., BALDWIN T., BOND F., COPESTAKE A. A. & FLICKINGER D. (2002). Multiword Expressions : A Pain in the Neck for NLP. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, p. 1–15, London, UK, UK : Springer-Verlag.

SAGOT B. (2010). The Lefff, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA).

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

SEDDAH D., TSARFATY R., KÜBLER S., CANDITO M., CHOI J. D., FARKAS R., FOSTER J., GOENAGA I., GOJENOLA GALLETEBEITIA K., GOLDBERG Y., GREEN S., HABASH N., KUHLMANN M., MAIER W., NIVRE J., PRZEPIÓRKOWSKI A., ROTH R., SEEKER W., VERSLEY Y., VINCZE V., WOLIŃSKI M., WRÓBLEWSKA A. & DE LA CLERGERIE E. V. (2013). Overview of the SPMRL 2013 Shared Task : A Cross-Framework Evaluation of Parsing Morphologically Rich Languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, p. 146–182, Seattle, Washington, USA : Association for Computational Linguistics.

SOCHER R., HUVAL B., MANNING C. D. & NG A. Y. (2012). Semantic Compositionality through Recursive Matrix-Vector Spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 1201–1211, Jeju Island, Korea : Association for Computational Linguistics.

STRATOS K. (2017). Reconstruction of Word Embeddings from Sub-Word Parameters. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, p. 130–135, Copenhagen, Denmark : Association for Computational Linguistics.

WANG S. & MANNING C. D. (2012). Baselines and Bigrams : Simple, Good Sentiment and Topic Classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 90–94, Jeju Island, Korea : Association for Computational Linguistics.

WIETING J., BANSAL M., GIMPEL K. & LIVESCU K. (2016). Charagram : Embedding Words and Sentences via Character n-grams. In *Proceedings of the 2016 Conference on Empirical Methods*

in Natural Language Processing, p. 1504–1515, Austin, Texas : Association for Computational Linguistics.

YU S., KULKARNI N., LEE H. & KIM J. (2017). Syllable-level Neural Language Model for Agglutinative Language. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, p. 92–96, Copenhagen, Denmark : Association for Computational Linguistics.