Robust-to-Noise Models in Natural Language Processing Tasks

Valentin Malykh
Neural Systems and Deep Learning Laboratory, Moscow Institute of Physics and Technology,
Samsung-PDMI Joint AI Center, Steklov Mathematical Institute at St. Petersburg
valentin.malykh@phystech.edu

Abstract

There are a lot of noisy texts surrounding a person in modern life. A traditional approach is to use spelling correction, yet the existing solutions are far from perfect. We propose a robust to noise word embeddings model which outperforms existing commonly used models like fasttext and word2vec in different tasks. In addition, we investigate the noise robustness of current models in different natural language processing tasks. We propose extensions for modern models in three downstream tasks, i.e. text classification, named entity recognition and aspect extraction, these extensions show improvement in noise robustness over existing solutions.

1 Introduction

The rapid growth of the usage of mobile electronic devices has increased the number of user input text issues such as typos. This happens because typing on a small screen and in transport (or while walking) is difficult, and people accidentally hit wrong keys more often than when using a standard keyboard. Spell-checking systems widely used in web services can handle this issue, but they can also make mistakes. These typos are considered to be noise in original text. Such noise is a widely known issue and to mitigate its presence there were developed spelling correcting systems, e.g. (Cucerzan and Brill, 2004). Although spelling correction systems have been developed for decades up to this day, their quality is still far from perfect, e.g. for the Russian language it is 85% (Sorokin, 2017). So we propose a new way to handle noise i.e. to make models themselves robust to noise.

This work is considering the main area of noise robustness in natural language processing and, in particular, in four related subareas which are described in corresponding sections. All the subareas share the same research questions applied to a particular downstream task:

RQ1. Are the existing state of the art models robust to noise?

RQ2. How to make these models more robust to noise?

In order to answer these RQs, we describe the commonly used approaches in a subarea of interest and specify their features which could improve or deteriorate the performance of these models. Then we define a methodology for testing existing models and proposed extensions. The methodology includes the experiment setup with quality measure and datasets on which the experiments should be run.

This work is organized as follows: in Section 2 the research on word embeddings is motivated and proposed, in further sections, i.e. 3, 4, 5, there are propositions to conduct research in the area of text classification, named entity recognition and aspect extraction respectively. In Section 6 we present preliminary conclusions and propose further research directions in the mentioned areas and other NLP areas.

2 Word Embeddings

Any text processing system is now impossible to imagine without word embeddings — vectors encode semantic and syntactic properties of individual words (Arora et al., 2016). However, to use these word vectors user input should be clean (i.e. free of misspellings), because a word vector model trained on clean data will not have misspelled versions of words. There are examples of models trained on noisy data (Li et al., 2017), but this approach does not fully solve the problem, because typos are unpredictable and a corpus cannot contain all possible incorrectly spelled versions of a word. Instead, we suggest that we should make

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 10–16
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

algorithms for word vector modelling robust to noise.

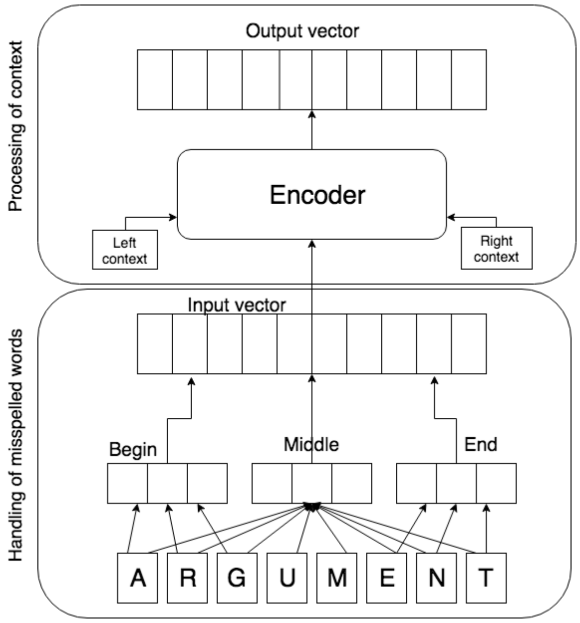

Figure 1: RoVe model architecture.

We suggest a new architecture **RoVe** (Robust Vectors).[1] It is presented on Fig. 1. The main feature of this model is open vocabulary. It encodes words as sequences of symbols. This enables the model to produce embeddings for out-of-vocabulary (OOV) words. The idea as such is not new, many other models use character-level embeddings (Ling et al., 2015) or encode the most common ngrams to assemble unknown words from them (Bojanowski et al., 2016). However, unlike analogous models, RoVe is specifically targeted at typos — it is invariant to swaps of symbols in a word. This property is ensured by the fact that each word is encoded as a bag of characters. At the same time, word prefixes and suffixes are encoded separately, which enables RoVe to produce meaningful embeddings for unseen word forms in morphologically rich languages. Notably, this is done without explicit morphological analysis. This mechanism is depicted on Fig. 2.

Another feature of RoVe is context dependency — in order to generate an embedding for a word one should encode its context (the top part of Fig. 1). The motivation for such architecture is the following. Our intuition is that when processing an OOV word our model should produce an embedding similar to that of some similar word from the training data. This behaviour is suitable for typos as well as unseen forms of known words. In the latter case we want a word to get an embedding similar to the embedding of its initial form. This process reminds lemmatisation (reduction of a word to its initial form). Lemmatisation is context-dependent since it often needs to resolve homonymy based on word's context. By making RoVe model context-dependent we enable it to do such implicit lemmatisation.

At the same time, it has been shown that embeddings which are generated considering word's context in a particular sentence are more informative and accurate, because a word's immediate context informs a model of the word's grammatical features (Peters et al., 2018). On the other hand, use of context-dependent representations allowed us to eliminate character-level embeddings. As a result, we do not need to train a model that converts a sequence of character-level embeddings to an embedding for a word, as it was done in (Ling et al., 2015).

2.1 Methodology

We suppose to compare RoVe with common word vector tools: word2vec (Mikolov et al., 2013) and fasttext (Bojanowski et al., 2016).

We score the performance of word vectors generated with RoVe and baseline models on three tasks: paraphrase detection, sentiment analysis, identification of text entailment. We consider these tasks to be binary classification ones, so we use ROC AUC measure for model quality evaluation.

For all tasks we suppose to train simple baseline models. This is done deliberately to make sure that the performance is largely defined by the quality of vectors that we use. For all the tasks we will compare word vectors generated by different modifications of RoVe with vectors produced by word2vec and fasttext models.

We presume to conduct the experiments on datasets for three languages: English (analytical language), Russian (synthetic fusional), and Turkish (synthetic agglutinative). Affixes have different structures and purposes in these types of languages, and in our experiments we show that our character-based representation is effective for all of them.

For the above mentioned tasks we are going to use the following corpora: Paraphraser.ru

[1] An open-source implementation is available here: https://gitlab.com/madrugado/robust-w2v

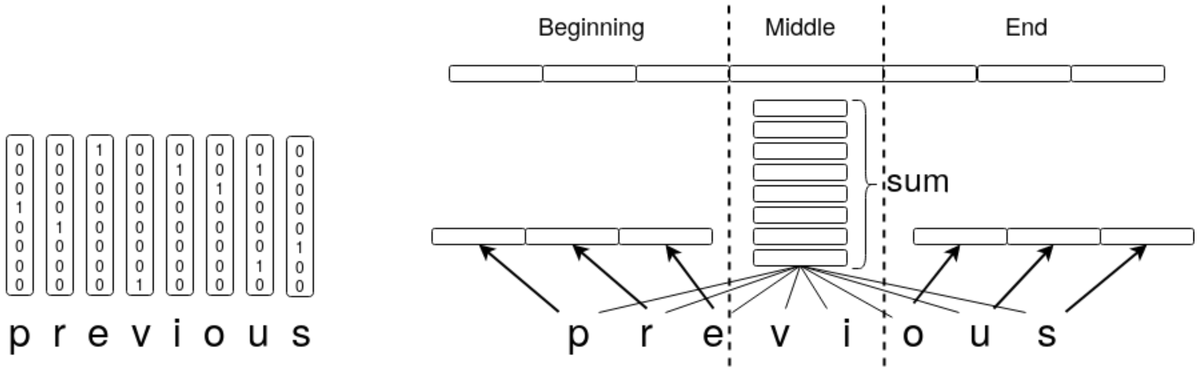

Figure 2: Generation of input embedding for the word *previous*. Left: generation of character-level one-hot vectors, right: generation of BME representation.

	English			Russian		
noise (%)	0	10	20	0	10	20
BASELINES						
word2vec	0.649	0.611	0.554	0.649	0.576	0.524
fasttext	**0.662**	0.615	0.524	0.703	0.625	0.524
RoVe						
stackedLSTM	0.621	0.593	0.586	0.690	0.632	0.584
SRU	0.627	0.590	0.568	0.712	0.680	0.598
biSRU	0.656	**0.621**	**0.598**	**0.721**	**0.699**	**0.621**

Table 1: Results of the sentiment analysis task in terms of ROC AUC.

(Pronoza et al., 2016) for the Russian language paraphrase identification task, Microsoft Research Paraphrase Corpus (Dolan et al., 2004) for the English language paraphrase identification task, Turkish Paraphrase Corpus (Demir et al., 2012) for the Turkish language paraphrase identification task; Russian Twitter Sentiment Corpus (Rubtsova, 2014) for the Russian language sentiment analysis task, Stanford Sentiment Treebank (Socher et al., 2013) for the English language sentiment analysis task; and Stanford Natural Language Inference (Bowman et al., 2015) for the English language natural language inference task.

2.2 Results

Due to lack of space we provide the results only for sentiment analysis task for the Russian and English languages and for natural language inference task for the English language.

There are three variants of the proposed RoVe model listed in Tables 1 and 2, these are ones using different recurrent neural networks for context encoding. The whole results are published in (hidden).

For both mentioned tables the robust word embedding model Rove shows better results for all noise level and both tasks, with the exception of zero noise for English language sentiment analysis task for which the fasttext word embeddings are showing better results. The latter could be explained as fasttext has been explicitly trained for this zero noise level, which is unnatural for human generated text.

	English		
noise (%)	0	10	20
BASELINES			
word2vec	0.624	0.593	0.574
fasttext	0.642	0.563	0.517
RoVe			
stackedLSTM	0.617	0.590	0.516
SRU	0.627	0.590	0.568
biSRU	**0.651**	**0.621**	**0.598**

Table 2: Results of the task on identification of textual entailment.

3 Text Classification

A lot of text classification applications like sentiment analysis or intent recognition are performed on user-generated data, where no correct spelling or grammar may be guaranteed.

Classical text vectorisation approach such as bag of words with one-hot or TF-IDF encoding encounters out-of-vocabulary problem given vast variety of spelling errors. Although there are successful applications to low-noise tasks on common datasets (Bojanowski et al., 2016; Howard and Ruder, 2018), not all models behave well with real-world data like comments or tweets.

3.1 Methodology

We do experiments on two corpora: Airline Twitter Sentiment [2] and Movie Review (Maas et al., 2011), which are marked up for sentiment analysis task.

We conduct three types of experiments: (a) the train- and testsets are spell-checked and artificial noise in inserted; (b) the train- and testsets are not changed (with the above mentioned exception for Russian corpus) and no artificial noise is added; and (c) the trainset is spell-checked and noised, the testset is unchanged.

These experimental setups are meant to demonstrate the robustness of tested architectures to artificial and natural noise.

As baselines we use architectures based on fasttext word embedding model (Bojanowski et al., 2016) and an architecture which follows (Kim et al., 2016). Another baseline, which is purely character-level, will be adopted from the work (Kim, 2014).

3.2 Results

Fig. 3 contains results for 4 models:

- FastText, which is recurrent neural network using fasttext word embeddings,
- CharCNN, which is a character-based convolutional neural network, based on work (Kim, 2014),
- CharCNN-WordRNN - a character-based convolutional neural network for word embeddings with recurrent neural network for entire text processing; it follows (Kim et al., 2016),
- and RoVe, which is a recurrent neural network using robust to noise word embeddings.

One could see in the figure that the model which uses robust word embeddings is more robust to noise itself starting from 0.075 (7.5%) noise level.

4 Named Entity Recognition

The field of named entity recognition (NER) received a lot of attention in past years. This task is an important part of dialog systems (Béchet, 2011). Nowadays dialog systems become more and more popular. Due to that the number of dialog system users is increased and also many users communicate with such systems in inconvenient environments, like being in transport. This makes a user to be less concentrated during a conversation and thus causes typos and grammatical errors. Considering this we need to pay more attention to NER models robustness to this type of noise.

4.1 Methodology

We conduct three types of experiments: (a) the trainset and testset are not changed and no artificial noise is induced; (b) the artificial noise is inserted into trainset and testset simultaneously; and (c) the trainset is being noised, the testset is unchanged.

These experimental setups are meant to demonstrate the robustness of tested architectures to artificial and natural noise (i.e. typos).

The proposed corpora to use are: English and Russian news corpora, CoNLL'03 (Tjong Kim Sang and De Meulder, 2003) and Persons-1000 (Mozharova and Loukachevitch, 2016) respectively, and French social media corpus CAp'2017 (Lopez et al., 2017).

We investigate variations of the state of the art architecture for Russian (Anh et al., 2017) and English (Lample et al., 2016) languages and apply the same architecture to the French language corpus.

5 Aspect Extraction

Aspect extraction task could provide information to make dialogue systems more engaging for user (Liu et al., 2010).

Therefore, we have decided to study the Attention-Based Aspect Extraction (ABAE) model (He et al., 2017) robustness using artificially generated noise. We propose three extensions for an ABAE model, which are supposedly more noise robust. There are:

- CharEmb - a convolutional neural network over characters in addition to word as a whole embeddings; these two embeddings are concatenated and used in ABAE model;
- FastText - an ABAE model using fasttext word embeddings;

[2] Publicly available here: https://www.kaggle.com/crowdflower/twitter-airline-sentimen

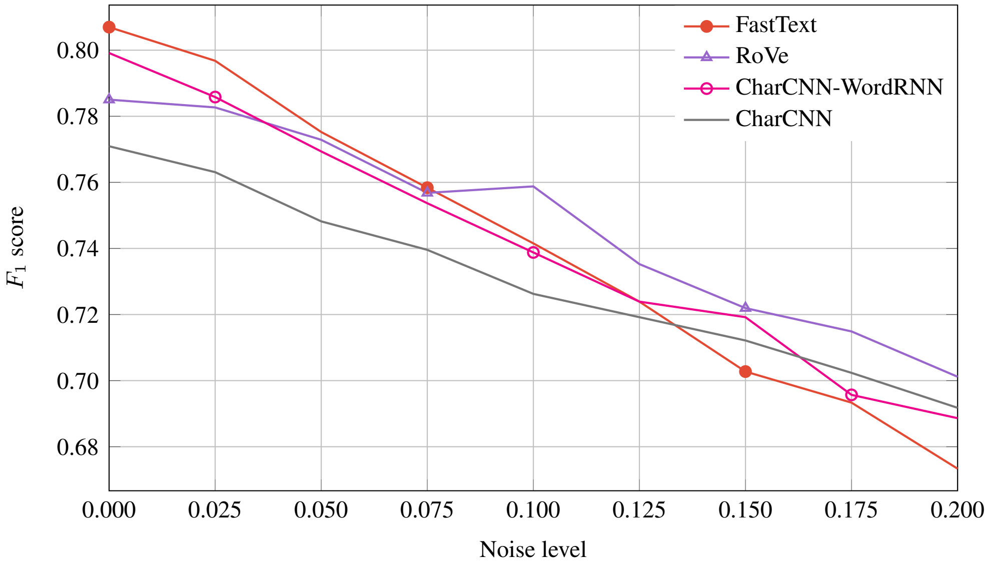

Figure 3: Airline Twitter Sentiment Dataset. Trained on spell-checked and noised data, tested on spell-checked and noised with the same noise level as the training set.

- RoVe - an ABAE model using robust word embeddings.

5.1 Methodology

As the noise model, we took simple character swapping with some probability, i.e. for any given string we go through it character by character and randomly decide if we need to replace this particular letter of the input with some random character.

As a quality measure we take F_1 (weighted by class representativity) score following (He et al., 2017). The authors of the original paper used data from the Citysearch corpus with user reviews on restaurants in New York city originally described in (Ganu et al., 2009). The reviews were labeled by human annotators with a set of categories, like “Food” or “Stuff”. The authors used only reviews with exactly one labeled category. So in the end a model predicts a label for a review in the unsupervised way. The label is considered to be the most probable aspect label.

5.2 Results

In Fig. 4 we show both the baseline ABAE model and its extended version proposed in this work. The original model has shown lower results for all lower noise levels, while all extensions show improvement over the original model. The RoVe extensions shows improvement for all noise levels over the original model and the other extensions. The full results for aspect extraction task are published in (Malykh and Khakhulin, 2018).

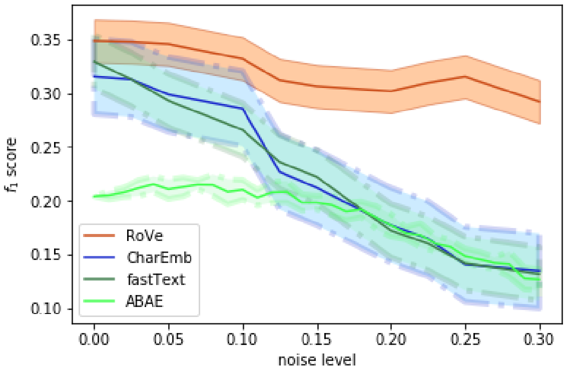

Figure 4: F_1 measure for ABAE model and proposed extensions.

6 Preliminary Results and Future Research Directions

In this work the research in four related subareas is proposed, these are word embeddings, text classification and named entity recognition and aspect extraction.

Preliminary experiments for the robust to noise word embeddings showed that explicit noise handling is better than implicit like in fasttext model. The preliminary results for the word embeddings had been published in (Malykh, 2017). The possible further research in that direction could be an investigation of embeddings for infix morphology languages, like Arabic and Hebrew.

In the downstream tasks experiments show that designed noise robustness improves quality on noisy data. For named entity recognition task the preliminary results are published in (Malykh and Lyalin, 2018), and for aspect extraction task the results are published in (Malykh and Khakhulin, 2018). The further research could be done in three directions. Firstly, all of the tasks could be applied to more languages. Secondly, for classification task corpora with more marked up classes could be used. This task is harder in general case, and there are some available corpora with dozens of classes. And last but not least, thirdly, the suggested methodology could be applied to the other subareas of natural language processing, like Automatic Speech Recognition and Optical Character Recognition, and achieve results in noise robustness improvement there.

References

Thanh L Anh, Mikhail Y Arkhipov, and Mikhail S Burtsev. 2017. Application of a hybrid bi-lstm-crf model to the task of russian named entity recognition. In *Conference on Artificial Intelligence and Natural Language*, pages 91–103. Springer.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. Linear algebraic structure of word senses, with applications to polysemy.

Frédéric Béchet. 2011. Named entity recognition. *Spoken Language Understanding: systems for extracting semantic information from speech*, pages 257–290.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christoper Manning. 2015. A large annotated corpus for learning natural language inference.

Silviu Cucerzan and Eric Brill. 2004. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Seniz Demir, Ilknur Durgar El-Kahlout, Erdem Unal, and Hamza Kaya. 2012. Turkish paraphrase corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: improving rating predictions using review text content. In *WebDB*, volume 9, pages 1–6. Citeseer.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 388–397.

Jeremy Howard and Sebastian Ruder. 2018. Fine-tuned language models for text classification. *arXiv preprint arXiv:1801.06146*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. pages 1746–1751.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Quanzhi Li, Sameena Shah, Xiaomo Liu, and Armineh Nourbakhsh. 2017. Data sets: Word embeddings learned from tweets and general data.

Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernández Astudillo, Silvio Amir, Chris Dyer, Alan W. Black, and Isabel Trancoso. 2015. Finding function in form: Compositional character models for open vocabulary word representation. *CoRR*, abs/1508.02096.

Jingjing Liu, Stephanie Seneff, and Victor Zue. 2010. Dialogue-oriented review summary generation for spoken dialogue recommendation systems. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 64–72. Association for Computational Linguistics.

Cédric Lopez, Ioannis Partalas, Georgios Balikas, Nadia Derbas, Amélie Martin, Coralie Reutenauer, Frédérique Segond, and Massih-Reza Amini. 2017. Cap 2017 challenge: Twitter named entity recognition. *arXiv preprint arXiv:1707.07568*.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics.

V. Malykh. 2017. Generalizable architecture for robust word vectors tested by noisy paraphrases. In *Supplementary Proceedings of the Sixth International Conference on Analysis of Images, Social Networks and Texts (AIST 2017)*, pages 111–121.

Valentin Malykh and Taras Khakhulin. 2018. Noise robustness in aspect extraction task. In *The Proceedings of the 2018 Ivannikov ISP RAS Open Conference*.

Valentin Malykh and Vladislav Lyalin. 2018. Named entity recognition in noisy domains. In *The Proceedings of the 2018 International Conference on Artificial Intelligence: Applications and Innovations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. 26.

Valerie Mozharova and Natalia Loukachevitch. 2016. Two-stage approach in russian named entity recognition. In *Intelligence, Social Media and Web (ISMW FRUCT), 2016 International FRUCT Conference on*, pages 1–6. IEEE.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Ekaterina Pronoza, Elena Yagunova, and Anton Pronoza. 2016. Construction of a russian paraphrase corpus: Unsupervised paraphrase extraction. 573:146–157.

Yuliya Rubtsova. 2014. Automatic term extraction for sentiment classification of dynamically updated text collections into three classes. In *International Conference on Knowledge Engineering and the Semantic Web*, pages 140–149. Springer.

R Socher, A Perelygin, J.Y. Wu, J Chuang, C.D. Manning, A.Y. Ng, and C Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. 1631:1631–1642.

Alexey Sorokin. 2017. Spelling correction for morphologically rich language: a case study of russian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 45–53.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.