Identification of Multiword Expressions: comparing the performance of a Conditional Random Fields model on corpora of written and spoken Italian

Ilaria Manfredi[1], Lorenzo Gregori[1]

[1]*University of Florence, P.zza San Marco 4, 50121 Florence, Italy*

Abstract

This paper describes an experiment that compares the performance of a Conditional Random Fields model on identification of Multiword expressions in corpora of spoken and written Italian. The model is trained on a corpus of spoken language and a corpus of written language annotated with Multiword expressions, then tested on two other corpora (one written and one spoken). This methodology provides very good results regarding Precision.

Keywords

Multiword Expressions, Conditional Random Fields, Spoken corpora

1. Introduction

"Multiword expression" (MWE) is a term used to refer to groups of words that display formal or functional idiosyncratic properties with respect to free word combinations, and therefore behave like a unit [1]. This notion encompasses a wide set of linguistic phenomena, of both semantic and syntactic nature, like idioms, verb-particle constructions, complex nominals, and support verb constructions. The computational treatment of MWEs notoriously poses a challenge in NLP [2], but in recent years a lot of effort has been put into the development of techniques and tools for the identification of MWEs in corpora. These are almost exclusively derived from, and tested on, written corpora. This leaves the study of MWEs in spoken varieties of languages, including Italian, a rather unexplored field.

Given the major differences between spoken and written language, we deemed it important to establish how an MWEs automatic extraction tool trained on written corpus performs on a spoken one, also considering the lack of specific resources for spoken corpora. We have decided to conduct an experiment training a Conditional Random Fields (CRF) model [3] to identify MWEs. The model was trained on both a corpus of spoken and one of written Italian; the two models obtained were then tested on corpora of spoken and written Italian, and their performances were evaluated. In § 2 we give an overview of existing research on MWEs and related resources for Italian; in § 3 we describe the resources used to build the training and test corpora; in § 4 we described the methodology followed to annotate the training corpora with MWEs and the testing; results of the experiment are presented in § 5 and discussed in § 6.

2. Related work

Identification of MWEs in corpora is essential for various NLP tasks such as machine translation and parsing, so a lot of research has been done on automatic acquisition of MWEs, both in general and for specific languages [4]. Many studies have explored the use of Association Measures for MWEs identification [5, 6, 7]; methodologies based on parallel corpora have also been investigated [8]. More recently, the use of different AI models has been tested for this task [9, 10]. Among these, CRF models has been used successfully in NLP for various sequence labeling tasks, including MWEs identification [11, 12, 13]. Given that, we have decided to use one of the CRF models available for our experiment (see § 4). As already mentioned all of these studies have been conducted on written corpora only, and so are the resources derived (mainly MWE annotated corpora and gold standard lists).

As for MWEs in spoken corpora, Strik et al. investigated possible ways of automatically identifying MWEs in Dutch speech corpora based on pronunciation characteristics; Trotta et al. built PoliSdict, a dictionary of Italian MWEs extracted from a corpus of political speech. To the best of our knowledge, this is the only resource of speech language MWEs existing for Italian. Other resources for Italian MWEs are PARSEME-It, a written corpus annotated with verbal MWEs [16, 17], and a validated dataset of MWEs from written corpora compiled by Masini et al. [19].

This brief overview highlights the gap in existing literature regarding MWEs from spoken language; hence, our experiment seeks to evaluate the performance of one

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

ilariamanfredi3@gmail.com (I. Manfredi); lorenzo.gregori@unifi.it (L. Gregori)

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

of the tools available, up to now tested only on written corpora.

3. Resources

For the experiment, we have used two training corpora and two test corpora (described in § 4.1) derived from the following resources.

KIParla [20] is a spoken corpus containing more than 112 hours of speech recorded in various settings from speakers of different areas of Italy, and is currently composed of two modules. The KIP module [21] contains speech of students and professors recorded in the Universities of Bologna and Turin.

IMAGACT is a corpus of approximately 1.8 million tokens[1] used for the creation of the IMAGACT Visual Ontology resource [22]; it contains texts of spoken Italian derived from LABLITA Corpus of Spontaneous Italian, LIP corpus, and the spoken section of CLIPS corpus. The materials contained are heterogeneous from a diaphasic, diastratic, and diatopic point of view (see Gagliardi for a detailed description).

CorDIC-scritto is a web corpus created within the RIDIRE project [24] containing written texts pertaining to five different semantic and functional domains: creative, bureaucratic, news, arts, economy[2].

PAISÀ [25] is a web corpus of approximately 250 million tokens containing documents from web pages. Part of the documents was obtained by retrieving pages using pairs words from the Italian basic vocabulary list as queries; others were derived from the Italian versions of various Wikimedia Foundation projects.

4. Methodology

This work has been conducted making use of the `mwetoolkit` software [26] for the extracting, filtering and annotating of the MWEs; the CRF model we have used is the one implemented in the CRFsuite software [27] and provided within the toolkit.

4.1. Training and test corpora

We have used the KIP module[3] of KIParla as the spoken training corpus and CorDIC-scritto as the written training corpus. As the spoken test corpus we have used IMAGACT. Lastly, for the written test corpus we have sampled PAISÀ to have approximately the same number of words of IMAGACT, in order for them to be comparable in size. Table 1 shows numbers of words and tokens of each of the corpora.

Table 1
Training and test corpora with number of words and tokens.

	Name	Words	Tokens
Spoken training	KIP	559,816	637,867
Written training	CorDIC	502,665	589,036
Spoken test	IMAGACT	1,366,305	1,870,272
Written test	PAISÀ	1,366,313	1,686,217

All of the corpora have been POS-tagged and lemmatized with Treetagger [28] using Baroni's parameter file [4].

4.2. Annotation of the training corpora

The first step to annotate the training corpora was the extraction of candidates, obtained by searching the corpora with sets of POS-patterns (see Ramish and Lenci et al. for an assessment of the method). The chosen POS-patterns were derived from the work of Masini et al., who provided a dataset of 1682 validated Italian MWEs extracted from written corpora with the POS-pattern method. We chose to use the top 20 POS-patterns in the dataset ranked by number of MWEs. Since the patterns in the dataset are provided according to the ISST-Tanl tagset[5], we first "translated" the tags to their respective ones in Baroni's tagset. The tagsets are not symmetrical (for example ISST-Tanl tags *RD* 'determinative article' and *RI* 'indeterminative article' are both *ART* 'article' in Baroni's tagset) so we computed again frequency of MWEs for each pattern and then took the top 20. The 20 POS-patterns used are bigrams and trigrams of adjectival, nominal, verbal, adverbial and prepositional patterns.

Using `mwetoolkit` functions, the corpora were searched and for every POS-pattern a list of candidates was obtained; each corpus was searched independently and the lists of candidates were examined separately. As a second step, all the lists of the candidates were filtered by number of occurrences: only candidates with a frequency of 4 or more were kept. Lists containing a high number of candidates were further filtered, before being manually examined: for KIP, lists having more than 150 candidates were ranked by LogLikelihood and the top 100 were examined; for CorDIC, lists with more than

[1] Here tokens are intended as single graphic units that include punctuation, symbols and words, as usual in computational linguistics

[2] See http://cordic.lablita.it/

[3] Compared to the original resource, available on https://kiparla.it/search/, our corpus lacks the documents BOC1006, BOD2008, TOA3005, TOD1005bis.

[4] https://home.sslmit.unibo.it/ baroni/collocazioni/itwac.tagset.txt.

[5] http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf

Table 2
Candidates extracted (f $>$ 3) and candidates examined for each POS-pattern.

POS-pat	cK	aK	cC	aC
A-N	189	100	263	100
PreArt-A-N	25	25	48	48
PreArt-N	729	100	1781	100
PreArt-N-Pre	37	37	240	100
N-A	258	100	697	100
N-PreArt-N	56	56	284	100
N-N	36	36	27	27
N-Pre-N	108	108	228	100
N-V	108	108	134	100
Pre-A-N	15	15	38	38
Pre-Art-N	115	115	255	100
Pre-DInd-N	15	15	28	28
Pre-N	664	100	1216	100
Pre-N-Pre	52	52	143	100
V-A	106	106	104	100
V-Adv	439	100	151	100
V-Art-N	148	148	69	69
V-PreArt-N	16	16	42	42
V-N	109	109	84	84
V-Pre-N	50	50	48	48
Total	3275	1496	5980	1584

100 candidates were ranked by LogLikelihood[6] and the top 100 were examined. In lists having less candidates than that, all of the candidates were examined. This way there is approximately the same number of candidates to be examined for each corpus: 1496 for KIP and 1584 for CorDIC.

Table 2 shows, for each POS-pattern, the number of candidates with frequency $>$ 3 in KIP (candK) and CorDIC (candC) and the number of candidates examined in each corpus (anK and anC). POS are abbreviated like this: A = adjective, N = noun, Pre-Art = articulated preposition, Pre = preposition, V = verb, Art= article, DInd = indefinite determiner, Adv = adverb.

As the final step, the remaining candidates from all the lists were manually examined. Candidates who showed some type of idiomaticity, fixedness, or were characterized by high familiarity of use were annotated as MWEs: in total, 214 MWEs for KIP and 204 for CORDIC. MWEs were tagged in their respective corpora using the IOB format [32]. In this process, attention has been put to only tag MWEs when they are in an idiomatic context, and not where they have a literal meaning.

Table 3
Occurrences of MWEs and Precision for each model on each corpus

	MWEs	Pr
S model IMAGACT	7508	0,974
S model PAISÀ	3337	0,908
W model IMAGACT	6291	0,978
W model PAISÀ	5047	0,946

4.3. Training and testing

The model was trained on MWE annotated KIP and CorDIC independently, using the functions of `mwetoolkit`; the training script was not modified and the features were kept as provided[7].

So we obtained two models, one trained on KIP (the 'spoken model') and one trained on CorDIC (the 'written model'). We used each of them to identify MWEs from IMAGACT and PAISÀ, with the aim to compare the results and determine if the best performance on spoken corpus comes from a spoken o written model, and vice versa.

5. Results

The spoken model tagged 7508 occurrences of MWEs in IMAGACT and 3337 in PAISÀ; the written model tagged 5047 occurrences of MWEs in PAISÀ and 6291 in IMAGACT. For a full evaluation of the models we need to compute Precision and Recall of the annotated corpora. Computation of Recall needs all the false negatives in test corpora to be identified; for that, we would need to manually annotate the entire corpora which is a very time-consuming task that requires multiple trained annotators. Another element of complexity for this task is to provide annotators with a precise definition of what to consider a MWE, as the distinction between MWEs and other types of word combinations is not always clear-cut. So, evaluation has been performed by manually computing Precision on a sample of 500 MWEs from each batch of results. Table 3 shows occurrences of MWEs and Precision at 500 for spoken and written models on each corpus.

6. Discussion

Results obtained show a great performance overall for both of the models, given the high value for Precision for all four of the corpora tagged. However, considering also the number of MWE occurrences tagged, we

[6] To calculate LogLikelihood for trigrams we have used the Ngram Statistics Package [30, 31]

[7] See https://gitlab.com/mwetoolkit/mwetoolkit3/-/blob/master/resources/default-config/listFeatures.txt

can see that the spoken model performed the worst on PAISÀ, having the lowest Precision and number of occurrences, while better results are achieved on the same corpus by the written model. On IMAGACT, both of the models performed very well, with the written model having the best Precision overall but slightly fewer occurrences of MWEs found. We have also counted the number of MWEs tagged (per lemmas) in IMAGACT, and how many of these were "new" compared to the ones annotated in the training corpora. The spoken model tagged 222 MWEs (per lemmas) of which 63 were new (28.4%) and the written model tagged 224 MWEs (per lemmas), 64 being new (28.6%), so the models performed similarly in this regard too. A slight difference in performance can be noted comparing Precision in tagging new MWEs: new MWEs found by spoken model account for a total of 119 occurrences, 46 of which results correctly tagged; new MWEs found by written model account for 123 occurrences, 60 of which are correctly tagged.

In conclusion, the results of this experiment show that on spoken corpora 'written models' perform similarly to 'spoken models'; this looks really promising, considering the lack of resources dedicated to MWEs in spoken language. Future works in this line of research include the computing of Recall for the models and qualitative evaluation of the MWEs extracted.

References

[1] F. Masini, Multi-word expressions and morphology, 2019. doi:10.1093/acrefore/9780199384655.013.611.

[2] I. A. Sag, T. Baldwin, F. Bond, A. A. Copestake, D. Flickinger, Multiword expressions: A pain in the neck for NLP, in: A. F. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing, Third International Conference, CICLing 2002, Mexico City, Mexico, February 17-23, 2002, Proceedings, volume 2276 of *Lecture Notes in Computer Science*, Springer, 2002, pp. 1–15. doi:10.1007/3-540-45715-1_1.

[3] J. D. Lafferty, A. McCallum, F. 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, Morgan Kaufmann Publishers Inc., San Francisco, California, 2001, p. 282–289.

[4] C. Ramish, A generic and open framework for multiword expression treatment: from acquisition to applications, Ph.D. thesis, Universitade Federal do Rio Grande do Sul, 2012.

[5] P. Pecina, A machine learning approach to multiword expression extraction, in: Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions, 2008, pp. 54–57.

[6] A. Fazly, S. Stevenson, Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures, in: Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 9–16. doi:10.3115/1613704.1613706.

[7] G. I. Lyse, G. Andersen, Collocations and statistical analysis of n-grams: Multiword expressions in newspaper text, in: Exploring newspaper language: Using the web to create and investigate a large corpus of modern Norwegian, 2012, pp. 79–110. doi:10.1075/scl.49.05lys.

[8] Y. Tsvetkov, S. Wintner, Extraction of multi-word expressions from small parallel corpora, Natural Language Engineering 18 (2012) 549–573. doi:10.1017/S1351324912000101.

[9] W. Gharbieh, V. Bhavsar, P. Cook, Deep learning models for multiword expression identification, in: Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 54–64. doi:10.18653/v1/S17-1006.

[10] R. Swaminathan, P. Cook, Token-level identification of multiword expressions using pre-trained multilingual language models, in: Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1–6. doi:10.18653/v1/2023.mwe-1.1.

[11] A. Maldonado, L. Han, E. Moreau, A. Alsulaimani, K. D. Chowdhury, C. Vogel, Q. Liu, Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking, in: Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), Association for Computational Linguistics, Valencia, Spain, 2017, pp. 114–120. doi:10.18653/v1/W17-1715.

[12] K. Nongmeikapam, D. Laishram, N. B. Singh, N. M. Chanu, S. Bandyopadhyay, Identification of reduplicated multiword expressions using crf, in: A. F. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 41–51. doi:10.1007/978-3-642-19400-9_4.

[13] M. Scholivet, C. Ramisch, Identification of ambiguous multiword expressions using sequence models and lexical resources, in: Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), Association for Computational Linguistics, Valencia, Spain, 2017, pp. 167–175. doi:10.18653/v1/

W17-1723.

[14] H. Strik, M. Hulsbosch, C. Cucchiarini, Analyzing and identifying multiword expressions in spoken language, Language Resources and Evaluation 44 (2010) 41–58. doi:10.1007/s10579-009-9095-y.

[15] D. Trotta, T. Albanese, M. Stingo, R. Guarasci, A. Elia, Multi-word expressions in spoken language: Polisdict, in: Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018: 10-12 December 2018, Accademia University Press, Torino, 2018. doi:10.4000/books.aaccademia.3654.

[16] J. Monti, M. P. Di Buono, F. Sangati, Parseme-it corpus: An annotated corpus of verbal multiword expressions in italian, in: Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017: 11-12 December 2017, Rome [online], Accademia University Press, 2017, pp. 228–233. doi:10.4000/books.aaccademia.2433.

[17] A. Savary, C. Ramisch, B. Guillaume, A. Hawwari, A. Walsh, A. Fotopoulou, A. Bielinskienė, A. Estarrona, A. Gatt, A. Butler, A. Rademaker, A. Maldonado, A. Villavicencio, A. Farrugia, A. Muscat, A. Gatt, A. Antić, A. De Santis, A. Raffone, A. Riccio, A. Pascucci, A. Gurrutxaga, A. Bhatia, A. Vaidya, A. Miral, B. QasemiZadeh, B. Priego Sanchez, B. Griciūtė, B. Erden, C. Parra Escartín, C. Herrero, C. Carlino, C. Pasquer, C. Liebeskind, C. Wang, C. Ben Khelil, C. Bonial, C. Somers, C. Aceta, C. Krstev, E. Bejček, E. Lindqvist, E. Erenmalm, E. Palka-Binkiewicz, E. Rimkute, E. Petterson, F. Cap, F. Hu, F. Sangati, G. Wick Pedro, G. Speranza, G. Jagfeld, G. Blagus, G. Berk, G. Attard, G. Eryiğit, G. Finnveden, H. Martínez Alonso, H. de Medeiros Caseli, H. Elyovich, H. Xu, H. Xiao, I. Miranda, I. Jaknić, I. El Maarouf, I. Aduriz, I. Gonzalez, I. Matas, I. Stoyanova, I.-P. Jazbec, J. Busuttil, J. Waszczuk, J. Findlay, J. Bonnici, J. Šnajder, J.-Y. Antoine, J. Foster, J. Chen, J. Nivre, J. Monti, J. McCrae, J. Kovalevskaitė, K. Jain, K. Simkó, K. Yu, K. Azzopardi, K. Adalı, L. Uria, L. Zilio, L. Boizou, L. van der Plas, L. Galea, M. Sarlak, M. Buljan, M. Cherchi, M. Tanti, M. P. Di Buono, M. Todorova, M. Candito, M. Constant, M. Shamsfard, M. Jiang, M. Boz, M. Spagnol, M. Onofrei, M. Li, M. Elbadrashiny, M. Diab, M.-M. Rizea, N. Hadj Mohamed, N. Theoxari, N. Schneider, N. Tabone, N. Ljubešić, O. Vale, P. Cook, P. Yan, P. Gantar, R. Ehren, R. Fabri, R. Ibrahim, R. Ramisch, R. Walles, R. Wilkens, R. Urizar, R. Sun, R. Malka, S. A. Galea, S. Stymne, S. Louizou, S. Hu, S. Taslimipoor, S. Ratori, S. Srivastava, S. R. Cordeiro, S. Krek, S. Liu, S. Zeng, S. Yu, Š. Arhar Holdt, S. Markantonatou, S. Papadelli, S. Leseva, T. Kuzman, T. Kavčič, T. Lynn, T. Lichte, T. Pickard, T. Dimitrova, T. Yih, T. Güngör, T. Dinç, U. Iñurrieta, V. Tajalli, V. Stefanova, V. Caruso, V. Puri, V. Foufi, V. Barbu Mititelu, V. Vincze, V. Kovács, V. Shukla, V. Giouli, X. Ge, Y. Ha-Cohen Kerner, Y. Öztürk, Y. Yarandi, Y. Parmentier, Y. Zhang, Y. Zhao, Z. Urešová, Z. Yirmibeşoğlu, Z. Qin, Stank, M. Cristescu, B.-M. Zgreabăn, E.-A. Bărbulescu, R. Stanković, PARSEME corpora annotated for verbal multiword expressions (version 1.3), 2023. URL: http://hdl.handle.net/11372/LRT-5124, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[18] F. Masini, M. Micheli, A. Zaninello, S. Castagnoli, M. Nissim, Multiword expressions we live by: A validated usage-based dataset from corpora of written italian, in: Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021, volume 2769, CEUR-WS.org, 2020. doi:10.4000/books.aaccademia.8710.

[19] F. Masini, M. S. Micheli, A. Zaninello, S. Castagnoli, M. Nissim, Mwe_combinet_release_1.0, 2020. URL: https://amsacta.unibo.it/id/eprint/6506/.

[20] C. Mauri, S. Ballarè, E. Goria, M. Cerruti, S. Francesco, Kiparla corpus: a new resource for spoken italian, in: Proceedings of the 6th Italian Conference on Computational Linguistics CLiC-it, CEUR-WS.org, 2019.

[21] E. Goria, C. Mauri, Il corpus kiparla: una nuova risorsa per lo studio dell'italiano parlato, in: CLUB Working Papers in Linguistics Volume 2, CLUB - Circolo Linguistico dell'Università di Bologna, 2018, pp. 96–116.

[22] M. Moneglia, S. Brown, F. Frontini, G. Gagliardi, F. Khan, M. Monachini, A. Panunzi, The IMAGACT visual ontology. an extendable multilingual infrastructure for the representation of lexical encoding of action, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 3425–3432.

[23] G. Gagliardi, Validazione dell'ontologia dell'azione IMAGACT per lo studio e la diagnosi del Mild Cognitive Impairment (MCI), Ph.D. thesis, Università degli Studi di Firenze, 2013.

[24] M. Moneglia, Il progetto ridire.it: un web corpus per l'accesso degli apprendenti l2 alla fraseologia italiana, in: Linguistica educativa: atti del XLIV Congresso internazionale di studi della Società di linguistica italiana (SLI): Viterbo, 27-29 settembre 2010, Bulzoni, 2012, pp. 411–423. doi:10.1400/202206.

[25] V. Lyding, E. Stemle, C. Borghetti, M. Brunello,

S. Castagnoli, F. Dell'Orletta, H. Dittmann, A. Lenci, V. Pirrelli, The PAISÀ corpus of Italian web texts, in: Proceedings of the 9th Web as Corpus Workshop (WaC-9), Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 36–43. doi:10.3115/v1/W14-0406.

[26] C. Ramisch, A. Villavicencio, C. Boitet, mwetoolkit: a framework for multiword expression identification, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010.

[27] N. Okazaki, Crfsuite: a fast implementation of conditional random fields (crfs), 2007. URL: http://www.chokkan.org/software/crfsuite/.

[28] H. Schmid, Probabilistic part-of-speech tagging using decision trees, in: Proceedings of the International Conference on New Methods in Language Processing, 1994.

[29] A. Lenci, F. Masini, M. Nissim, S. Castagnoli, G. Lebani, L. Passaro, M. Senaldi, How to harvest word combinations from corpora: Methods, evaluation and perspectives, Studi e saggi linguistici 55 (2017) 45–68.

[30] S. Banerjee, T. Pedersen, The design, implementation, and use of the ngram statistics package, in: Computational Linguistics and Intelligent Text Processing, volume 2000, 2003, pp. 370–381. doi:10.1007/3-540-36456-0_38.

[31] T. Pedersen, S. Banerjee, B. McInnes, S. Kohli, M. Joshi, Y. Liu, The ngram statistics package (text::NSP) : A flexible tool for identifying ngrams, collocations, and word associations, in: Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, Association for Computational Linguistics, Portland, Oregon, 2011, pp. 131–133.

[32] L. Ramshaw, M. Marcus, Text chunking using transformation-based learning, in: Third Workshop on Very Large Corpora, 1995.