SinNer@Clef-Hipe2020 : Sinful adaptation of SotA models for Named Entity Recognition in French and German

Pedro Javier Ortiz Suárez, Yoann Dupont, Gaël Lejeune, Tian Tian

SinNer@Clef-Hipe2020 : Sinful adaptation of SotA models for Named Entity Recognition in French and German

Pedro Javier Ortiz Suárez[1,3], Yoann Dupont[2], Gaël Lejeune[2], and Tian Tian[2]

[1] ALMAnaCH, Inria, Paris, France
[2] STIH, Sorbonne Université, Paris, France
[3] Sorbonne Université, Paris, France
[1]{firstname.lastname}@inria.fr
[2]{firstname.lastname}@sorbonne-universite.fr

Abstract. In this article we present the approaches developed by the Sorbonne-INRIA for NER (SinNer) team for the CLEF-HIPE 2020 challenge on Named Entity Processing on old newspapers. The challenge proposed various tasks for three languages, among them we focused on Named Entity Recognition in French and German texts. The best system we proposed ranked third for these two languages, it uses FastText embeddings and Elmo language models (FrELMo and German ELMo). We show that combining several word representations enhances the quality of the results for all NE types and that the segmentation in sentences has an important impact on the results.

Keywords: Named Entity Recognition · Historical Texts · German · French · ELMo · CRFs. · Sentence Segmentation

1 Introduction

Among the aspects for which Natural Language Processing (NLP) can be useful for Digital Humanities (DH) figures prominently Named Entity Recognition. This tasks interests researchers for numerous reasons since the application can be pretty wide. Among other usages we can cite genealogy or history for which finding mention of persons and places in texts is very useful. Researchers in digital literature have shown a great interest in Named Entities since it can help for instance to highlight the path of different characters in a book or in a book series. There can be cross-fertilization between NER and DH since some researchers showed that some particular properties of literature can help to build better NER systems [1]. Apart from literature, NER can also be used more generally to help refine queries to assist browsing in newspaper collections [17]. Like other NLP tasks, NER quality will suffer from different problems related to variations in the input data : variation in languages (multilinguality), variation in the quality of the input data (OCR errors mainly) and specificity of the application domain (literature VS epidemic surveillance for instance). These difficulties can be connected with the challenges for low level NLP tasks highlighted by Dale *et al.* [2]. In this CLEF-HIPE challenge [5], the variation in language and in text quality will be the main problems even if the specificity of the application can be of great interest.

NER in old documents represent an interesting challenge for NLP since it is usually necessary to process documents that show different kind of variations as compared to the particular laboratory conditions on which NER systems are trained. Most NER systems are usually designed to process clean data. On the other hand, there is the multilingual issue since NER systems have been designed primarily for English, with assumptions on the availability of data on one hand and on the universal nature of some linguistic properties on the other hand.

The fact that the texts processed in Digital Humanities are usually not born-digital is very important since, even after OCR post-correction, it is very likely that some noise would be found in the text. Other difficulties will arise as well in those type of documents. The variation in language is one of them since contemporary English will clearly not be the most frequent language. It is interesting for researchers to check how much diachronic variation has an influence on NER systems [4]. It makes it even more important to work on multilingual NER and to build architecture that need less training data [23]. More generally, NER in ancient texts represents a great opportunity for NLP to compare to main approaches to handle variation in texts : adapting the texts to an existing architecture via modernization or normalization [14] or adapting the pipeline to non standard data (OCR noise, language variants ...) via domain adaptation or data augmentation techniques [8].

In Section 2 we will present a brief state-of-the-art for Named Entity Recognition with a focus on digitized documents. Section 3 and 4 will respectively devoted to the description of the dataset of CLEF-HIPE 2020 shared task [5] and the methods we developed to extract NE for French and German. The results of our systems will be described in Section 5 and in Section 6 we will give some conclusions and perspectives for this work.

2 State of the Art for Named Entity Recognition

Named Entity Recognition came into light as a prerequisite for designing robust Information Extraction (IE) systems in the MUC conferences [9]. This task soon began to be treated independently from IE since it can serve multiple purposes, like Information retrieval or Media Monitoring for instance[31]. As such, shared task specifically dedicated to NER started to rise like CONLL [29]. Two main path were followed by the community: (I) since NER was at first used for general purposes, domain extension start to gain interest [6]; (II) since the majority of NER systems were designed for English, the extension to novel languages (including low resource languages) became of importance [25].

Regarding approaches, one can say that NER followed the different fashions in NLP. The first approaches were based on dictionaries, gazeeters and hand-crafted rules. Initially NER was considered to be solved by a patient process involving careful syntactic analysis [10]. Supervised learning approaches came to fashion with the increase of available data and specifically the rise of shared tasks on the subject. Decision trees and Markov models were soon outperformed with Condition Random Fields. By taking advantage of the sequentiality of textual data, CRF helped to set new state-of-the-art results in the domain [7]. Since supervised learning results were bound by the size of training data, lighter approaches were tested in the beginning of the 2000's, we can cite weakly supervision [30] and active learning [26].

During a time, most of promising approaches involved an addition to improve CRfs: embeddings [20], (bi-)LSTMs [12] [15] or contextual embeddings [21]. More recently, the improvements in contextual word embeddings made the CRFs models disappear of the architecture reaching state-of-the-art results, see [27] for a review on the subject and a very interesting discussion on the limits attained by the state-of-the-art systems.

3 Dataset for the CLEF-HIPE shared task

The provided corpus in French and German both contain training data (train) and development (dev) data whereas, for English only development data was provided. We chose to work only on French and german The table 1 shows some statistics of this dataset. The size of the train dataset was two time higher for French than for German wheras the development sets had roughly the same size. As usual in NER, persons (Pers) and locations (Loc) are the most frequent entity types.

	Tokens	Documents	Segments	Labeled named entities				
				Pers	Loc	Org	Time	Prod
Train Fr	166217	158	19183	3067	2513	833	273	198
Dev Fr	37592	43	4423	771	677	158	69	48
Train De	86960	104	10353	1747	1170	358	118	112
Dev De	36175	40	4186	664	428	172	73	53

Table 1. Statistics of training and dev data in French and German

The dataset is represented in CONLL format with one token per line. Table 2 shows an extirp of the train dataset in French. For each document, general information were provided. Newspaper and date may have been features useful for recognising entities but we dit not take advantage of it. Each document was composed of segments, starting with "# segment ..." corresponding to lines in the original documents. Each segment was tokenized to correpsond to the CONLL format with one token per line. These two notions, segments and tokens, are very important since they do not always match the type of unit usually processed in NLP. Segments seldom correspond to sentences so that there is a need to concatenate the segments to get the raw text and then segment it into sentences. This is very interesting since it gets us close to real-world conditions rather than laboratory conditions, and we show in Section 5.2 that this segment VS sentence question has an important influence on the results. Regarding tokens, the tokenization is obviously not perfect. We can see that there are non-standard words and bad tokenization due to the OCR output (in red in Table 2). If we concataenate the tokens we get the sequence "Su. _sss allemands" instead of "Suisse allemande". These non-standard words make the Named Entity Recognition task more complicated and, again, more realistic.

# language = fr										
# newspaper = EXP										
# date = 1918-04-22										
# document_id = EXP-1918-04-22-a-i0077										
# segment_iiif_link = https://iiif.dhlab.epfl.ch/iiif_impresso/\dots										
Lettre	O	O	O		O	O	O	_	_	_
de	O	O	O		O	O	O	_	_	_
la	O	O	O		O	O	O	_	_	_
Su	B-loc	O	B-loc.adm.reg		O	O	B-loc.adm.nat	Q689055	_	NoSpaceAfter
.	I-loc	O	I-loc.adm.reg	O	O	I-loc.adm.nat	Q689055	_	_	
_	I-loc	O	I-loc.adm.reg	O	O	I-loc.adm.nat	Q689055	_	NoSpaceAfter	
sss	I-loc	O	I-loc.adm.reg	O	O	I-loc.adm.nat	Q689055	_	_	
allemands	I-loc	O	I-loc.adm.reg		O	O	O	Q689055	_	EndOfLine
# segment_iiif_link = https://iiif.dhlab.epfl.ch/iiif_impresso/\dots										
(	O	O	O		O	O	O	_	_	NoSpaceAfter
Nous	O	O	O		O	O	O	_	_	_
serons	O	O	O		O	O	O	_	_	_
heureux	O	O	O		O	O	O	_	_	_
de	O	O	O		O	O	O	_	_	_
publier	O	O	O		O	O	O	_	_	_
de	O	O	O		O	O	O	_	_	_
temps	O	O	O		O	O	O	_	_	_
à	O	O	O		O	O	O	_	_	EndOfLine

Table 2. Example extracted from French training dataset

4 CRFs and Contextualized Word Embeddings for NER

4.1 CRF model

SEM[4] [3] is a free NLP tool that relies on linear-chain CRFs [11] to perform tagging. SEM uses Wapiti [13] v1.5.0[5] as linear-chain CRFs implementation. SEM uses the following features for NER:

- token, prefix/suffix from 1 to 5 and a Boolean isDigit features in a [-2, 2] window;
- previous/next common noun in sentence;
- 10 gazetteers (including NE lists and trigger words for NEs) applied with some priority rules in a [-2, 2] window;

[4] available at : https://github.com/YoannDupont/SEM
[5] available at : https://github.com/Jekub/Wapiti

– a "fill-in-the-gaps" gazetteers feature where tokens not found in any gazetteer are replaced by their POS, as described in [24]. This features used token unigrams and token bigrams in a [-2, 2] a window.
– tag unigrams and bigrams.

We trained a CLEF HIPE specific model by optimizing L1 and L2 penalties on the development set. The metric used to estimate convergence of the model is the error on the development set ($1 - accuracy$). For French, our optimal L1 and L2 penalties were 0.5 and 0.0001 respectively (default Wapiti parameters). For German, our optimal L1 and L2 penalties were 1.0 and 0.0001 respectively.

One interest of SEM is that it has a builtin sentence tokenizer for french using a rule-based approach. By default, CLEF-HIPE documents segments are only newlines. As a result, some NE mentions span across multiple segments, making it very hard to identify them correctly. It is to be expected that models trained (and labelling on) sentences would yield better performances than those trained (and labelling on) segments. SEM makes it simple to switch between different sequence segmentations, which allowed us to label sentences and output segments. SEM's sentence segmentation engine works using mainly local rules at token level rule to determine whether a token is the last of a sequence. It also uses non-local rules to remember whether a token is between parentheses or french quotes to not segment automatically within them. Since we work at token level, we had to adapt some rules to fit CLEF-HIPE tokenization. For example, SEM decides at tokenization stage whether a dot is a strong punctuation or part of a larger token, as for abreviations. This has the advantage of making sentence segmentation easier. CLEF-HIPE tokenization systematically separates dots, so we adapted some tokenization rules to apply them at sentence segmentation stage. We decided to not consider a dot as a sentence terminator if the previous token was in a lexica of titles or functions.

Another interest is that SEM has an NE mention broadcasting process. Mentions found at least once in a document are used as a gazetteer to tag unlabeled mentions within said document. When a new mention overlaps and is strictly longer than an already found mention, the new mention will replace the previous one in the document.

4.2 Contextualized word embeddings

Embeddings from Language Models (ELMo) [22] is a Language Model, i.e, a model that given a sequence of N tokens, $(t_1, t_2, ..., t_N)$, computes the probability of the sequence by modeling the probability of token t_k given the history $(t_1, ..., t_{k-1})$:

$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k \mid t_1, t_2, \ldots, t_{k-1}).$$

However, ELMo in particular uses a bidirectional language model (biLM) consisting of L LSTM layers, that is, it combines both a forward and a backward language model jointly maximizing the log likelihood of the forward and backward directions; ELMo also computes a context-independent token representation via token embeddings or via a CNN over characters.

When included in a downstream model, as it is the case in this paper, ELMo collapses all L layers into a single vector, generally computing a task specific weighting of all biLM layers applying layer normalization to each biLM layer before weighting. Following [22], we use in this paper ELMo models where $L = 2$, i.e., the ELMo architecture involves a character-level CNN layer followed by a 2-layer biLSTM.

4.3 ELMo-LSTM-CRF

The LSTM-CRF is a model originally proposed by Lample et al. [12] is a BiLSTM pre-appended by both character level word embeddings and pre-trained word embeddings and pos-appended by a CRF decoder layer. For our experiments, we follow the same approach as Ortiz Suárez et al. [18] by using the Bi-LSTM-CRF implementation of [28] which is open source and readily available[6] and is designed to easily pre-append contextualized word-embeddings to the model.

Historically English had receivend the most attention in NER, with some recent developments in German, Dutch and Spanish by [28]

And additional term of comparison was identified in a recently released state-of-the-art language model for French, CamemBERT [16], based on the RoBERTa architecture pre-trained on the French sub-corpus of the newly available multi-lingual corpus OSCAR [19].

5 Results and Discussion

5.1 Official shared-task results

Results of our 3 runs compared to the best run on the NERC-coarse shared-task for french and german are given in table 3. For each of those tasks, we are the third best ranking team. We only did very minimal adaptation of existing systems. The most notable one was to use custom sentence segmentation instead of given segments for french and using some language-specific lexica for our CRF model in german. Other than that, we only optimized hyper-parameters on the dev set. This clearly illustrates the power of contextual embeddings and today's neural network architectures. This simple adaptation to the task allowed us to be among the three best ranking teams. This is encouraging in terms of usability of SotA models on real-world data.

5.2 Study of sequence segmentation

As can be seen in table 4, sentence segmentation allows to improve results by 3.5 F1 points. This is due to the fact that some entities were split across multiple segments in the original data. Using a custom sentence segmentation allows to

[6] Available at: `https://github.com/ufal/acl2019_nested_ner`.

RUN	FRENCH			GERMAN		
	P	R	F1	P	R	F1
winner	83.1	84.9	84	79	80.5	79.7
run 1	77.8	79.4	78.6	63.1	66.6	64.8
run 2	78.8	80.2	79.5	65.8	65.8	65.8
run 3	70.2	57.9	63.5	64.4	43.8	52.1
average	70.2	66.7	67.6	63.8	58.1	60.0
median	71.5	68.6	68.6	66.8	57.7	64.5

Table 3. Strict results for our systems compared to the winning system

have entities in a single sequence. This benefits the models both at training and at label level, where systems can access a more proper context. The cost of using another segmentation is relatively cheap, as SEM can process nearly 1GB of raw text per hour.

A per entity comparison is also available in table 4. One can see that the improvement of sentence segmentation is not very significant for locations (Loc). It is due to two facts : (I) locations are usually small in number of tokens and therefore less prone to be sperated in two segments and (II) there was less room from improvement since they were the easiest entity type to detect (86.35% F1-measure). To the contrary, products (Prod), usually longer in tokens, were very hard to predict with only 48.57% F1-measure and benefited the most from segmentation in sentences (+16 percentage points in F1-measure).

ENTITY	P		R		F1	
	segments	sentences	segments	sentences	segments	sentences
Loc	85.21	87.73 (+2.52)	87.52	87.08 (-0.44)	86.35	87.41 (+1.06)
Org	70.62	71.33 (+0.71)	62.78	65.64 (+2.86)	66.47	68.37 (+1.90)
Pers	80.24	84.64 (+4.40)	76.88	82.09 (+5.21)	78.52	83.35 (+4.83)
Prod	62.96	75.86 (+12.90)	39.53	56.41 (+16.88)	48.57	64.71 (+16.14)
Time	86.21	90.91 (+4.70)	78.12	87.72 (+9.60)	81.97	89.29 (+7.32)
Global	81.03	84.46 (+3.43)	81.61	84.46 (+2.85)	79.52	83.01 (+3.49)

Table 4. Detailed comparison between segments and sentences on dev dataset

6 Conclusion

In this article we presented three methods developed for the Named Entity Recognition task in French and German ancient texts. The first method relied on linear-chain CRFs while the other two methods use a Bidirectional LSTM and a bidirectional Language Model (ELMo). The later outperformed the CRF model and achieved rank 3 on the NER task in both French and German. We also showed that the type of sequences used has a significant influence on the results. When we segment in sentences rather than using the segments of the dataset as it is the results are systematically much better, with an exception for locations where the gain is marginal. This proves that sentence segmentation remains a key component of efficient NLP architectures, in particular for models taking advantage of the context.

As a future work it would be interesting to assess the importance of noise in the data. For instance, by comparing the results of NER on texts obtained via different OCR tools. The influence of the qualitative jumps in the data, which is common in Digital Humanities, is an important aspect to evaluate the robustness of the system in real-world conditions rather than laboratory conditions.

References

1. Brooke, J., Hammond, A., Baldwin, T.: Bootstrapped text-level named entity recognition for literature. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 344–350. Association for Computational Linguistics, Berlin, Germany (Aug 2016)
2. Dale, R., Somers, H.L., Moisl, H.: Handbook of Natural Language Processing. Marcel Dekker, Inc., USA (2000)
3. Dupont, Y.: Exploration de traits pour la reconnaissance d'entités nommées du français par apprentissage automatique. In: 24e Conf'erence sur le Traitement Automatique des Langues Naturelles (TALN). p. 42 (2017)
4. Ehrmann, M., Colavizza, G., Rochat, Y., Kaplan, F.: Diachronic evaluation of ner systems on old newspapers. Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016) pp. 97–107 (2016)
5. Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)
6. Evans, R.: A framework for named entity recognition in the open domain. In: In Proceedings of the Recent Advances in Natural Language Processing (RANLP). pp. 137–144 (2003)
7. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. p. 363–370. ACL '05, Association for Computational Linguistics, USA (2005)
8. Ghannay, S., Caubrière, A., Estève, Y., Laurent, A., Morin, E.: End-to-end named entity extraction from speech (2018)
9. Grishman, R., Sundheim, B.: Design of the muc-6 evaluation. In: Proceedings of the 6th Conference on Message Understanding. p. 1–11. MUC6 '95, Association for Computational Linguistics, USA (1995)
10. Hobbs, J.R.: The generic information extraction system. In: Proceedings of the 5th Conference on Message Understanding. p. 87–91. MUC5 '93, Association for Computational Linguistics, USA (1993)

11. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning (ICML) 2001, Williams College, Williamstown, MA, USA. pp. 282–289 (2001)
12. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 260–270. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/N16-1030, `https://www.aclweb.org/anthology/N16-1030`
13. Lavergne, T., Cappé, O., Yvon, F.: Practical very large scale CRFs. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 504–513. Association for Computational Linguistics (2010)
14. Leaman, R., Lu, Z.: TaggerOne: joint named entity recognition and normalization with semi-Markov Models. Bioinformatics **32**(18), 2839–2846 (06 2016)
15. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1064–1074. Association for Computational Linguistics, Berlin, Germany (Aug 2016)
16. Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., Sagot, B.: CamemBERT: a tasty French language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7203–7219. Association for Computational Linguistics, Online (Jul 2020), `https://www.aclweb.org/anthology/2020.acl-main.645`
17. Neudecker, C., Wilms, L., Faber, W.J., van Veen, T.: Large-scale refinement of digital historic newspapers with named entity recognition. In: Proceedings of IFLA 2014 (2014)
18. Ortiz Suárez, P.J., Dupont, Y., Muller, B., Romary, L., Sagot, B.: Establishing a new state-of-the-art for French named entity recognition. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 4631–4638. European Language Resources Association, Marseille, France (May 2020)
19. Ortiz Suárez, P.J., Romary, L., Sagot, B.: A monolingual approach to contextualized word embeddings for mid-resource languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1703–1714. Association for Computational Linguistics, Online (Jul 2020), `https://www.aclweb.org/anthology/2020.acl-main.156`
20. Passos, A., Kumar, V., McCallum, A.: Lexicon infused phrase embeddings for named entity resolution. In: Proceedings of the Eighteenth Conference on Computational Natural Language Learning. pp. 78–86. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2014)
21. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018)
22. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018)

23. Rahimi, A., Li, Y., Cohn, T.: Massively multilingual transfer for NER. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 151–164. Association for Computational Linguistics, Florence, Italy (Jul 2019)
24. Raymond, C., Fayolle, J.: Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. In: TALN'10 (2010)
25. Rössler, M.: Adapting an NER-system for German to the biomedical domain. In: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP). pp. 95–98. COLING, Geneva, Switzerland (Aug 28th and 29th 2004)
26. Shen, D., Zhang, J., Su, J., Zhou, G., Tan, C.L.: Multi-criteria-based active learning for named entity recognition. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04). pp. 589–596. Barcelona, Spain (Jul 2004)
27. Stanislawek, T., Wróblewska, A., Wójcicka, A., Ziembicki, D., Biecek, P.: Named entity recognition - is there a glass ceiling? In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). pp. 624–633. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
28. Straková, J., Straka, M., Hajic, J.: Neural architectures for nested NER through linearization. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. pp. 5326–5331 (2019)
29. Tjong Kim Sang, E.F., De Meulder, F.: 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. p. 142–147. CONLL '03, Association for Computational Linguistics, USA (2003)
30. Yangarber, R.: Counter-training in discovery of semantic patterns. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. pp. 343–350. Association for Computational Linguistics, Sapporo, Japan (Jul 2003)
31. Yangarber, R., Lin, W., Grishman, R.: Unsupervised learning of generalized names. In: In Proceedings of the International Conference on Computational Linguistics (ICCL). pp. 1–7 (2002)