Review of the Ph.D. manuscript "Named Entity Recognition" by Ing. Michal Konkol

Christophe Cerisara

February 22, 2016

1 Objectives and achievements

In this thesis, the author identifies two main issues that are related to classical Named Entity Recognition systems:

- The necessity to fine-tune and manually adapt most NER systems to every new domain and language
- The lack of external knowledge used in these systems.

Based on these observations, he then proposes to design a complete Named Entity Recognition system that:

- Gives improved performance over the state-of-the-art for Czech, and also works well for a few other European languages;
- Is relatively generic and domain-agnostic thanks to semi-supervised training and the use of unlabeled corpora;
- Further addresses the Named Entity Disambiguation challenge;
- Is reusable and supports standard interfaces.

The work achieved in this thesis clearly fulfills in great parts all of these four objectives. Hence, the developed system is shown to compare favorably with other state-of-the-art systems, especially on the Czech language. This in particular results from the introduction of semantic features, which further improve the genericity of the system by leveraging large unlabeled corpora that may be independent of any given domain. Moreover, some interesting and encouraging preliminary results are given on a simplified Named Entity Disambiguation task, and a version of the software has been integrated into the standard GATE platform for Natural Language Processing.

Of course, with regard to the two motivating issues identified in the introduction, the outcome is much more mitigated, which is expected given the scope of these issues. First, although exploiting unsupervised corpora indeed improves genericity, the final system is not completely satisfying from this point of view: Previous works on NER with lexical semantic embeddings have already been proposed and have performed succesfully on several tasks and domains (see references later in this review). Applying such semantic features on several languages is more original, and multilinguality is thus an important aspect of this thesis. Similarly, the type of external knowledge used in this work does not really depart from the state-of-the-art.

But I would like to insist on the fact that the four main objectives have clearly been fulfilled, and I consider that the contributions and results presented in this thesis are successful and important. Of course, every aspect of the document is not always perfect: I thus review next the document chapter by chapter and then propose a global conlusion in the end. Although I may appear to often focus on negative aspects of the work, because these are the few aspects that depart from an otherwise global work of high quality, I insist on the fact that the following review does not jeopardize the overall quality of the work, which is of high standard.

2 Thesis report and litterature review

- Chap1 nicely motivates Named Entity Recognition.

- Chap2 is nicely written and very pedagogical. It clearly defines the standard evaluation metrics used in the domain.

- Chap3 describes the related works: it is very good and pleasant to read. On page 14, there is a sentence about NER authors who often use the term unsupervised for semi-supervised system: I think this statement should be taken with caution, because the terminology about weakly supervised approaches used in every domain is never very clear nor consensual in my view; this is why very often such terms are defined at the beginning of every paper that uses them. Furthermore, I don't believe it is useful to really discuss this lexical issue; the actual amount and type of supervision in every work is more important to consider. Hence, on page 15, although some bootstrapped systems use a few number of initial annotated exemples, there is still a large difference between such systems and others that exploit thousands of annotated plus millions of unannotated data, and are still called semi-supervised. In addition, there may be much stronger supervision in purely 'unsupervised' systems (e.g. in the model topology or other types of resources) than with a few labeled instances.

  More importantly, I regret the lack of references in the neural networks and deep learning domains: such works do exist in NER, and have already proven to give very good results. They are further very promising for the near future, and so I think it is important to at least cite and discuss them. Hence, on page 15, it is said that CRF are considered as the best method. I'm afraid the situation is not so clear nowadays with deep networks. For instance, the best results I know of have been obtained with a bidirectional-LSTM-CRF and very good results have also been reported with other deep networks. But apart from this omission, I would like to point out the very good and nice overview of existing NER systems, especially in Section 3.3.

- Chap4 describes some of the most common models used for NER, but there are other more advanced and complex architectures that are not described, especially based on neural-networks. It is of course not possible to review all of them in the thesis, but it would be better to at least clearly states that this overview does not cover the whole domain, and give some pointers to other types of models. The second part of the chapter with an overview of features is much better.

- Chap5 presents a selection of semantic vector space models. I think an important category is missing here: words embeddings, such as Glove, word2vec, or Collobert's embeddings, which all have been shown to carry very strong semantic information. On page 40, the discussion about local vs. global context, or more generally about the type of relations exhibited by various semantic models is very interesting and could have been longer, with deeper studies and references, because interpretability is still an open and important question for the community.

- Chap6 is one of the core contributions, about semantic features in NER. I review it next.

- Chap7 presents a study on the variations of the standard segmentation approach used in sequence tagging tasks. This contribution is very nice, and very well done ! Many test cases and situations are discussed and studied in a very rigorous way. Note however that both on pages 53 and 60, a limited set of results is given because of 'space constraints'. This may be fine in an article, but is not acceptable in a thesis report. So please give all results and extensively comment them.

- Chap8 presents nice experiments on morphological features. Note that a few sentences about the experimental setup (in Section 8.2) are repeated several times from the previous chapters and it may have been better to have a single basic experimental setup description with much more details, for instance in Section 6.1.2.

- Chap9 gives premiminary results on a simplified Named Entity Disambiguation task. The choice of a limited task, where only entity forms are given in the knowledge base, could have been further motivated and justified. It is also not very clear what is the precise contribution of the author with regard to the creation of the corpus, which could also have been described in more details: who did the annotations ? How many annotators

? What about inter-annotator Kappa ? The review of various string distance is quite good, but nothing is said about their respective advantages or use cases. In 9.3, why do you use both similarity and dissimilarity, which are totally correlated ? Isn't it a misuse of parameters in the model ? In Figure 9.3, could you explain why don't all curves converge to 75% for threshold=1 ?

- Chap10 describes the practical design of the system, and chapter 11 concludes the thesis. Note that in the proposed future work, the first proposed direction of research is about training a semantic model for the specific task of NER. This is where the author should definitely look at neural networks, because this is exactly what is done in every neural net with embeddings: initializing the embeddings with a generic semantic space, and then fine-tuning them for the task at hand.

Globally, the thesis is very well written, self-contained and very pedagogical: nearly every fundamental aspects of NER is justified and clearly described, making the thesis a useful reading for every new researcher in the field. This is a positive aspect of the thesis, but it also has its counterpart, that is experts in the field may prefer to skip all these basics and have a more lengthy explanation of the advanced original contributions. I know this is a delicate balance, compromise to find when writing a thesis, between a good description of the fundamentals and a thorough study of the contributions, and just like any compromise, it is impossible to satisfy everybody. Also, the contributions only cover 38 pages, from p.45 to p.83, which make the document a bit unbalanced towards non-contribution sections, with 44 pages. So a good solution also may be to make the thesis longer so that it includes everything required, and even more.

Finally, the thesis document is globally very good, but may be a bit too much structured as a concatenation of articles. It could be made better by further gluing all ideas together and giving a more general view of all contributions of this work.

3 Scientific contributions, publications

The author's publications are quite correct, with mainly TSD papers, 2 Workshops, 2 Lecture Notes, and one paper in an IF=2.24 journal. Did the author participate to the SemEval'2014 challenge, and if so, what were his results ?

This thesis reconsiders all the design of a complete NER systems, by studying every successive part of a good NER system: segment representations (this chapter is especially well designed), which features to use, and in particular morphology and semantic features (main contribution), which classifier to use (ME or CRF) although this point is probably the less well treated, as deep neural networks are completely missing, disambiguation, evaluation metrics and named entity disambiguation. Revisiting every such stage of a NER system is not new by itself and has already been realized a number of times in previous works, but this study is made in a quite rigorous way and with a constant multi-lingual point of view, focused on Czech, English but also several other important European languages, which is not so common in the litterature.

An important contribution is given in chapter 6. The first experiment in Section 6.1.1 is not really compliant with the objective of the thesis, that is building a NER system without too much manual efforts, because the groups are created by hand. Sometimes, the experimental setup could be described in greater details, especially with regard to how cross-validation is performed: is the atomic item a speaker, a sentence or a token ? This is important because cross-validation assumes independence between the train and test parts, which may be problematic at the token level. Also, Section 6.2.2 about the "unsmooth LDA" is not very clear, because vanilla LDA exploits a Dirichlet prior, and is thus smooth by default. So what does unsmooth LDA mean ?

Concerning the experimental methodology, it is brilliant in some parts, for instance in Section 7, but may create some doubts elsewhere: for instance, in Section 6.2.6, several systems with a different number of parameters are compared. This may not be a problem, because the number of parameters is not the only factor to take into account when comparing systems, but it would have been good to discuss this factor, compare the number of parameters and check the optimum as a function of the number of parameters for each method.

The neural-based systems are missing both in the related works and in the comparison with the state-of-the-art, although they have been proposed as soon as 2011. Hence, the paper from Collobert "NLP almost from scratch" gives with a deep neural network an F1 of 88.68% on CoNLL'03 with the unlabeled Reuters corpus and 89.59% with gazettes. The 2015 paper by Huang et al. "Bidirectional LSTM-CRF Models for Sequence Tagging" also gives 90.10%. I appreciate the fact that in this thesis, results as good as these ones, and even slightly better, are obtained: this is impressive ! However, a nice aspect of these two neural-based works is their strong genericity: indeed, both results are obtained with a task- and domain-independent model that is applied to several other NLP tasks than NER with the same set of features in both papers. This is particularly interesting with regard to the objectives of this thesis, which is also concerned with building a generic system. So this type of works should definitely be considered and discussed in the thesis.

More generally, I'm a bit frustrated by the lack of discussion and results about out-of-vocabulary words, which is I believe an important issue for NER: How does you system handle new unknown person/city names ?

Also, your system is presented like a mostly fully automatic one that uses as few manual rules as possible; but what about dates, urls, zip codes ? Don't you use any manual rule for them ? Aren't they country-specific ?

The results on NED are interesting and promising, but seem rather preliminary, because the target task is a simplified NED task, and so it can not be compared with the state-of-the-art. There is actually no such attempt in the thesis at making such a comparison.

Finally, despite these few details, the overall quality and importance of the proposed contributions is high, thanks in particular to the rigorous and precise investigations of fundamental aspects of NER systems, and also to the good multi-lingual aspects and results of the work.

4 Recommendations

Given the previous remarks, I recommend the thesis for defence.

CERISARA Christophe

Cerisara

CHARLES UNIVERSITY PRAGUE
faculty of mathematics and physics

Západočeská universita v Plzni
Děkanát FAV
Ing. Jaroslav Toninger
Univerzitní 8
Plzeň
306 15

Review of the doctoral thesis of Ing. Michal Konkol: Named Entity Recognition

Thesis Content

The topic of the thesis is recognition of named entities, which are generally multiword sequences representing various objects like people, organizations, places, etc. Named entity recognition is a well-studied problem of natural language processing and is often used as a preprocessing step during many NLP tasks.

The NER task and various evaluation metrics are described in Chapter 2. Related work, existing corpora and current state-of-the-art NER systems are presented in Chapter 3. Machine learning algorithms used in successful NER systems (HMM, SVM, ME, MEMM and CRF) and features used by those systems are sketched in Chapter 4. The Chapter 5 concludes the summary part of the thesis and introduces latent semantic models used later in the thesis.

The main parts of the thesis are the next four chapters, which describe research carried out by the author. The Chapter 6 describes two NER systems employing latent semantics. The second one is the most interesting NER system described in the thesis. It is language independent (i.e., requires no language-specific rules nor a lemmatizer) and reaches near state-of-the-art performance on English, Czech and Spanish and state-of-the-art performance on Dutch.

The effect of various lemmatization techniques on Czech NER is described in Chapter 8, resulting in a Czech-specific NER system with slightly higher performance than the language independent system described in Chapter 6.

Influence of various segment representations on NER system is evaluated in Chapter 7. Several segment representations are compared on four corpora in different languages, using paired Student's t-test with two confidence levels.

Chapter 9 describes entity linking task, which is to link named entity occurrences to real-world entities in a given knowledge base. Only a specific subtask of named entity linking is discussed in the chapter, with the knowledge base being a list of people names without additional information. Performance of various string similarity metrics on this subtask is evaluated using a hand-made corpus.

Chapter 10 lists a few implementation details of (publicly unavailable, to my best knowledge) created NER system and fulfillment of thesis goals is discussed in Chapter 11.

Thesis Contributions

According to my opinion, the most significant contribution of the thesis is the language independent system described in Chapter 6. The system extensively evaluates the effect of several set of features – stemming, LDA and word clusters created using 5 semantic spaces – on four corpora. Such detailed evaluation of individual features across multiple languages help other researchers to select features during development of new NER systems. The results are extensively compared to state-of-the-art in Table 6.5 and Subsection 6.2.10. However, this comparison is the weakest part of this chapter:

- Two systems better than the cited state-of-the-art for English are not mentioned, namely (Tkachenko and Simanovsky 2012, Named Entity Recognition: Exploring Features) reaching 91.02% F1-score and (Luo et al. 2015, Joint named entity recognition and disambiguation) reaching 91.2% F1-Score. The latter one may be too new, but (Tkachenko and Simanovski 2012) is mentioned in the thesis in Chapter 3, so it should definitely be included.
- System better than the cited state-of-the-art for Czech is not mentioned, namely (Demir and Özgür 2014, Improving Named Entity Recognition for Morphologically Rich Languages using Word Embeddings) reaching 75.61% F1-score, although it is mentioned in Chapter 3.
- For Czech, no comparison with (Straková et al. 2013, A new state-of-the-art Czech named entity recognizer) is performed.

The last issue deserves more details. Performance of Czech NER systems is measured on Czech Named Entity Corpus (CNEC), which differs from many other corpora by using embedded entities and rich two-level hierarchy. Several systems developed by the authors of the CNEC or their colleagues, notably (Ševčíková et al. 2007, Zpracování pojmenovaných entit v českých textech), (Kravalová et al. 2009, Czech Named Entity Corpus and SVM-based Recognizer) and (Straková et al. 2013, A New State-of-The-Art Czech Named Entity Recognizer) all used the same evaluation metric (F1-score on entities including the embedded ones, but only classes from the first round of annotations). The evaluation script has been published in 2013 and has since been part of CNEC release. The author of this thesis used the script to evaluate the NER system (Konkol 2013, CRF-based Czech Named Entity Recognizer and Consolidation of Czech NER Research) and achieved 79% F1-score, compared to 82.82% achieved by (Straková et al. 2013).

However, in the cited paper (Konkol 2013, CRF-based Czech Named Entity Recognizer and Consolidation of Czech NER Research) the author create a variant of the CNEC corpus (with no embedded entities and using coarser hierarchy) and evaluate results only on this variant since, namely in (Konkol et al. 2014, Named entity recognition for highly inflectional languages: Effect of various lemmatization and stemming approaches), (Konkol et al. 2014, Latent semantics in named entity recognition), (Konkol et al. 2015, Segment representations in named entity recognition) and in this thesis. This decision of not comparing with (Straková et al. 2013), which was state-of-the-art in 2013 (and probably still is), is highly questionable. It is also problematic to say in (Konkol et al. 2014, Named entity recognition for highly inflectional languages: Effect of various lemmatization and stemming approaches), that "These results outperform the current state-of-the-art on the CNEC 1.1 corpus", because they do so only using metric chosen by the authors of the paper, and most likely would not be state-of-the-art when using metric chosen by the CNEC authors.

That being said about shortcomings of Chapter 6, we now return to the rest of the thesis. Another considerable and self-contained contribution are the significance tests of various segment representations performed in Chapter 7. As mentioned by the authors, many researchers adopted the BILOU representation following the (Ratinov and Roth 2009), and this representation is shown to perform suboptimally on several datasets. However, the credibility of the significance tests is reduced by the fact that quite a low performing baseline is used in the experiments – for example, 83.47% F1-score for English BILOU compared to 91.02% from (Tkachenko and Simanovsky 2012), or 69.21% F1-score for Czech BILOU compared to 75.61% from (Demir and Özgür 2014).

The effect of various lemmatizers on NER system evaluated in Chapter 8 forms additional contribution of the thesis. The comparison includes most used lemmatizers/stemmers for Czech and interestingly evaluate the effect of these tools on real natural language processing pipeline. The best system described in this chapter is the best Czech NER system described in the thesis. Unfortunately, it is not again compared to (Straková et al. 2013).

The named entity linking experiments from Chapter 9 have also their merit. Notably, the created corpus linking personal names to the list of known people can be used by others when evaluating the performance of their named entity linking systems.

Technical Comments

- Usually, textual description of a NER system is missing many details, so for replicability, it would be best if the described NER systems were available for download. This seems to be more and more common in latest publications.
- Sections 4.1.1 and 4.1.2: The sections about HMM and SVM are particularly not well written. For example, I do not understand what you mean by "The Baum-Welch algorithm can be used to improve the parameters of HMM using unmarked texts" – Baum-Welch is used to train HMM using annotated texts (i.e. text with NER classes). In the description of SVM, "line" before (4.1) should be "hyperplane", equation (4.4) should contain "b", and SVM works also for not linearly separable data (both by using hinge loss and nonlinear kernels), which should be at least mentioned.

- Page 28: The best presented result for Czech NER was definitely not achieved with SVM.
- Page 30: Usually the described disadvantage of MEMM is the "label bias problem" (Lafferty et al, 2001, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data), not data sparseness.
- Page 31, page 68: General statements like "CRF are considered to be the most successful method for NER" should be supported by some references. For example, I do not consider CRF to be the most successful method for NER, given results of systems based on artificial neural networks like (Collobert et al. 2011, Natural Language Processing (almost) from Scratch), (Demir and Özgür 2014, Improving Named Entity Recognition for Morphologically Rich Languages using Word Embeddings), (Chiu and Nichols 2015, Named Entity Recognition with Bidirectional LSTM-CNNs) or (Ling et al. 2015, Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation).
- Page 33: Gazetteers are not defined yet, so either the term should be explained or at least a reference to Section 4.2.4 provided.
- Page 38: Segment representation used by corpus annotation and segment representation used by a NER system are quite different things – talking about BIO representation of CoNLL-200[23] corpora seem inappropriate in this context. For corpus annotation, any scheme able to capture multiword entities can be used without changing the results, while for a NER system the segment representation changes its performance. Also it does not make sense to use BILOU scheme for corpus annotation, while it does make sense for a NER system. (However, in your terminology, I believe CoNLL-2003 datasets use BIO-1, not BIO-2 as you suggest.)
- Page 40: Multiletter variables should not be typeset as `$Dirichlet$` or `$Multinomial$` which disrupt spacing, but using for example `$\mathit{Dirichlet}$`.
- Page 50: You should probably say "In this chapter", not "In this paper".
- Page 50: BIO representation would be better than BIO "format".
- Table 6.3: It would be probably more interesting to show results for cluster sizes 50, 100, 200 and 500, because for 1000 and 5000 the results are substantially worse.
- Figures 9.1 and 9.2: The values for 2 and more surface forms are not very readable. It would be better to either scale the *y* axis so that the maximum value would be the value for 2 surface forms and the value for 1 surface form would be written above its "overflowing" bar, or to use logarithmic scale.
- Page 75: I believe Levenshtein distance is defined as "the minimal number of insertions, deletions and replacements to make two strings equal", not as (9.1). The equation (9.1) is only one possible formulas for computing it.
- Page 86: You say that the whole pipeline is available through the GATE plugin system, but I could not find it anywhere. If you mean that the pipeline is publicly available, please provide a link or reference; if the pipeline is not publicly available, you should say something like "the whole pipeline is implemented as GATE plugins".

Language Comments

The thesis is written using good English with several minor language issues. Overall, the thesis reads well.

- Page 25: Sentence starts with “hidden”.
- Page 26, 50, 55: “*will* be described”, “we *will* firstly introduce”, “we *will* follow with a description” – use present tense instead of future tense.
- Page 32: “property of *the* feature is *a* language dependency” should be “property of a feature is language dependency”.
- Page 32: “to a shortened form, which has the *meaning* of a multiple characters of this type” should be written as for example “which represents”.
- Page 33: “or *the* n-gram can be used” – drop the definite article.
- Page 33: “Stemming is a task which is *trying* to” – do not use continuous tense.
- Page 36: “Wikipedia is a rich source *on* information”.
- Page 36: “The results *are obviously dependent* on the language of Wikipedia”.
- Page 43: “*There is an assumption* that these contexts...” – use something like “We assume that these contexts...”.
- Page 49: “An approximative clustering method *have* to be used”.
- Page 55: “morphological information that is lost *at* stemming.”.
- Page 55: “approaches *for* stemming and lemmatization.”
- Page 68: “Keep in mind, that *in* this way, we do not ...”.
- Page 81: “The input of a module is *the* text and annotations...”, “the output of a module *are*”.

Conclusion

The author demonstrated the ability to carry out independent research and his thesis contains several original results in the area of Named Entity Recognition. I therefore recommend Michal Konkol's Ph.D. thesis for defense.

Sincerely,

RNDr. Milan Straka., Ph.D.
Institute of Formal and Applied Linguistics
Malostranské nám. 25
118 00 Praha 1
Czech Republic

Review of doctoral thesis “Named Entity Recognition” of Michal Konkol by Hristo Tanev, Ph.D., Joint Research Centre, European Commission

The thesis Named Entity Recognition of engineer Michal Konkol is dedicated to the important problems of named entity recognition and disambiguation. The author uses machine learning to identify named entities in texts. The main contribution of this thesis to the state-of-the-art are the experiments with different semantic and morphological features for machine learning as well as feature combinations. The thesis also studies the application of different string similarity metrics for detection of named entity variants.

The author formulates four main goals of his thesis:

1. Develop new recognition methods and features to improve performance for Czech and other languages.

2. Propose semi -supervised approaches to improve the adaptability of NER

3. Experiment with disambiguation on small subset of selected named entities.

4. Create quality and reusable NER system, which will provide standard interfaces

The thesis is focused mostly on the Czech language and its peculiarities from the point of view of the named entity recognition task. The author successfully experiments also with English, Spanish, German and Dutch, thus proving the multilingual nature of the presented methods for feature selection. The content of the thesis is based on ten previous scientific publications of engineer Konkol, which shows the significance of the presented ideas.

The thesis describes the task and the thesis goals in the first and second section, then it presents a comprehensive overview of the related work in the third section. Then, it proceeds with description of the exploited machine learning approaches and features in the following sections. The work on the thesis resulted in the development of a statistical named entity recognition system, based on the GATE framework. The design of this system is briefly described in the thesis. All the four goals of thesis have been accomplished.

The thesis contributes to the state-of-the-art in Czech named entity recognition. The proposed feature combinations result in about 10% performance improvement with respect to the best existing named entity recognition system for Czech.

In my opinion, the thesis by Michal Konkol fulfils all the conditions for gaining the PhD. Degree.

18/02/2016

Ispra, Italy

Hristo Tanev, Ph.D