Government Domain Named Entity Recognition for South African Languages

Roald Eiselen

Centre for Text Technology, North-West University, Potchefstroom Campus, South Africa
Roald.Eiselen@nwu.ac.za

Abstract

This paper describes the named entity language resources developed as part of a development project for the South African languages. The development efforts focused on creating protocols and annotated data sets with at least 15,000 annotated named entity tokens for ten of the official South African languages. The description of the protocols and annotated data sets provide an overview of the problems encountered during the annotation of the data sets. Based on these annotated data sets, CRF named entity recognition systems are developed that leverage existing linguistic resources. The newly created named entity recognisers are evaluated, with *F*-scores of between 0.64 and 0.77, and error analysis is performed to identify possible avenues for improving the quality of the systems.

Keywords: Language resource development, South African languages, named entity recognition

1 Introduction

Named entity recognition (NER) is the process of automatically classifying different unique identifiers, named entities (NE), according to a predefined set of types. The automatic recognition of named entities is an important, well-understood technology, with a number of scientific events targeting NER (Doddington et al., 2004; Grishman & Sundheim, 1996a, 1996b; Nadeau & Sekine, 2007; Tjong Kim Sang & De Meulder, 2003) and a number of publications on the required data sets, features and algorithms that can be used for the development of NER. Even though extensive work has been done internationally, relatively little work has been done on developing the resources required for creating NE recognisers in the South African languages, apart from Afrikaans (Fourie et al., 2014; Matthew, 2013; Puttkammer, 2006). The research presented here provides an overview of language resources developed for ten South African languages for government domain NER.

Named entity recognition can be solved in a number of different ways using rule-based, supervised, semi-supervised or unsupervised methods (Nadeau & Sekine, 2007). Supervised methods typically perform best, but the development of annotated data sets is a time-consuming and training-intensive undertaking, which is why supervised NER systems for most of the South African languages have not been developed to date.

This paper describes the process and results of the effort to develop annotated named entity data sets for ten[1] of the official languages of South Africa. The first part of the paper describes the data annotation process, with specific focus on the annotation protocols and challenges experienced during the annotation process. The next section provides a short description of the automatic NER systems developed for each language. Finally we provide results for the quality of the NER systems and show *F*-score results of between 0.64 & 0.77 for the respective languages, with disjunctive languages generally performing better than the conjunctive languages. All of the resources developed as part of this project are available from the Resource Management Agency [2] under the Creative Commons Attribution Licenses.

2 Background

Over the past decade the South African Department of Arts and Culture (DAC) has funded a variety of projects with the aim of developing textual language resources and human language technologies for the official languages of South Africa. The NER work described here is part of the NCHLT Text Phase II project, which was completed at the end of 2015. The project aimed to annotate at least 15,000 named entity tokens for each of the official languages of South Africa, and based on these annotations develop supervised automatic NER systems for each language.

South Africa has eleven official languages belonging to four language families: The conjunctively written Nguni languages - isiZulu, isiXhosa, isiNdebele, and SiSwati; the disjunctively written Sotho languages - Setswana, Sesotho, Sesotho sa Leboa and Tshivenda; the disjunctively written Tswa-Ronga language Xitsonga; and the Germanic languages Afrikaans and English. The project to develop NER resources for these languages is part of this effort and an extension of a previous project that produced parallel data sets annotated for lemmatisation, morphology, and part of speech; monolingual text corpora, as well as lemmatisers, morphological decomposers and part of speech taggers (Eiselen & Puttkammer, 2014). This second phase of the project leveraged the resources developed during the previous project, either in the form of data, or as feature generators for the automatic NER systems.

The project was coordinated by the North-West University's Centre for Text Technology, but included language experts and annotators from six other research and development institutions. Each of the relevant resources is described in detail in the following sections.

3 Language Resources

3.1 Protocols and Annotated Data

During the initial stages of the project, the main focus was the development of the annotated data sets for each of the languages. Named entities consist of three main categories (with sub-categories) to be recognised, *viz.* entity names (specifically *person*, *location* and *organisation*), temporal

[1] English is excluded since most of these resources are already available for English.

[2] http://rma.nwu.ac.za/

expressions (specifically *date*, *time*, and *duration*), and number expressions (specifically *money*, *measure*, *percentage*, and *cardinal number*). For the purpose of this research, the annotation of named entities is limited to entity names, since these are the most useful types for other systems and technologies that can leverage NER, such as information extraction, sentiment analysis, text anonymisers, and machine translation (Babych & Hartley, 2003; Grishman, 2003; Kushida et al., 2012; Turchi et al., 2012). The annotation scheme and definitions for each of the entity types are based on the Conference of Natural Language Learning's shared tasks (Tjong Kim Sang & De Meulder, 2003) that also focused on entity names, rather than the other two categories. According to their scheme each token is assigned to one of four categories:

1. PER – person names;
2. ORG – organisations;
3. LOC – locations; and
4. OUT – any token that doesn't belong to one of the other categories.

Since we didn't want to lose an opportunity to annotate information to the resources developed in this project, we added a fifth category, miscellaneous (MISC), for items that are in one of the other two categories, temporal and number expressions, as well as obvious entities that do not fit into the above categories, such as publications, laws and language names. These items can then be annotated into additional subcategories in a future project.

In addition to the basic scheme and definitions discussed above, the project followed the following annotation principles in guiding decisions about whether a token or tokens form an NE:

1. The token must be a unique identifier;
2. The token must be a proper name, most likely written with capital letters;
3. The name of the entity must be assigned through some official process such as birth certification, official registration or assignment through a law or governmental agency; and
4. In metonymical expressions where the exact type is unclear, the most prototypical interpretation should be assigned.

The most problematic aspects of identifying NEs according to the protocols as defined by Tjong Kim Sang and De Meulder (2003) and the above principles is that there are a number of edge cases that are extremely difficult to categorise. Some examples of these occur in a phrase like "I am going to IBM" where IBM could be either marked as ORG or LOC. Furthermore, the conjunctive languages, along with Afrikaans, very often inflect a named entity to form an identifier that is not unique anymore, but is still a designator. Examples such as "South Africans" or "the South African coastline" where the NE South Africa, is no longer a unique identifier. In such cases, annotators were instructed to annotate these items as MISC since they have a degree of uniqueness, but doesn't meet the requirements set out above.

All data was annotated according to the established Inside, Outside, Beginning (IOB) labelling scheme (Ramshaw & Marcus, 1999), that is ideal for training sequence labellers. In order to facilitate easier and more accurate annotations, an extension of a previously developed annotation environment, LARA3 (CTexT, 2015), was developed, which facilitates simple and structured annotation.

One of the main aims of the DAC development of human language technologies is to assist the South African government in generating documentation in the native languages of its citizens and making data more broadly available to the South African community. With this aim in mind, the NER resources developed in this project focused exclusively on governmental data, either from websites, government digests or legislation. The implication of this is that the nature of the entities found in the data will likely be very different from various other domains, and various item types, such as laws and the official designation of government employees, are very prevalent in the data. Consistently identifying and annotating these types of items turned out to be especially difficult, and also had an impact on the automatic NER systems developed on the basis of these annotated data sets.

3.2 Automatic Named Entity Recognition Systems

In the second part of the NCHLT Text Phase II project, automatic NER systems for each of the languages were developed to form, in most cases, baseline systems that could be used for other development projects, or as starting points from which to improve NER systems for the South African languages. Although several different techniques have been shown to be accurate NER classification approaches, it was decided to use linear-chain conditional random fields (CRFs) with L2 regularisation, since this has been shown to be an effective and scalable technique to solve sequence labelling problems in the NER domain (Das & Garain, 2014; Konkol & Konopík, 2013; McCallum & Li, 2003). The CRFs compiled for these languages are based on and compatible with the open source CRF++ implementation (Kudo, 2005).

The baseline feature set used for all languages was typical NER features based on the graphemic features of a token, as set out in Table 1, all of which are added as binary features. Although these are very widely used features, the internal capitalisation, rather than just mixed case, is important, since the conjunctive languages often use a lower-case morpheme prior to the designator, which is a good indicator that a token belongs to an NE, for example *kaCarnegie* or *neDAC*.

In addition to the graphemic features, the models also included gazetteer information, with specific gazetteers for names, surnames, locations, and organisations.

Unfortunately, there are limited gazetteers available for most of the South African languages, with Afrikaans being the only exception for which good location, person name and surname gazetteers are available. During the first part of the NCHLT project tokens in the 50,000 word annotated corpus were marked as NEs, however they were not classified as specific NE types. From these sets, a restricted list of NEs were generated to be used as a feature in the identification of the NEs. One mitigating factor with regard to the gazetteers, however, is that many of the indigenous languages make use of loan words with a language-specific prefix when referring to locations, for example *i-India* is used to refer to "India" in isiZulu and *waseSomalia* refers to "Somalia" in SiSwati. These types of "loaned" English NEs meant that English language gazetteers could be used with a form of substring matching that verified the affixes associated with a capitalised word, and matching the remaining string to an English gazetteer. In addition to the gazetteers and graphemic features, the NER system also used the previously developed part of speech taggers as a feature for NER.

Feature	Examples
Capitalisation	Initial capitalisation ALL CAPS MixedCase Internal capitalisation (e.g. eThekwini)
Punctuation	Starts or ends with punctuation Has internal punctuation
Digits	Digit pattern Roman numeral Words with digits Date and year formats

Table 1: Grapheme features

The conjunctive languages shared a couple of problematic issues beyond the availability of gazetteers. Firstly, the inflectional nature of the language means that gazetteers have limited coverage since many of the forms in the inflectional paradigm will not be present in gazetteers. However, most location and organisation named entities in the conjunctive languages require a nominal prefix that is followed by a capitalised character, e.g. *waseNingizimu Afrika* (*South Africa*). Although these constructs are good indicators of an NE, the distinction between the different classes is not as easy, since many of these nominal prefixes are shared across two or three of the classes. Even with this problem, adding features indicating that a prefix followed by capitalisation is possible for a particular class did improve the conjunctive NER system by between 0.03 and 0.05 as measured by *F*-score.

4 Evaluation

4.1 Experimental Design

Named entity recognisers are generally evaluated with three different evaluation metrics, namely Precision, Recall and *F*-score, although there are different approaches to calculating these metrics, depending on how partial or overlapping entities are treated (Grishman & Sundheim, 1996a; Sundheim, 1996; Tjong Kim Sang & De Meulder, 2003). In the following evaluations we follow the simplest, but strictest definition of correct entities as described in the CoNLL 2002 and 2003 shared tasks. According to this approach, a correctly classified entity is any entity that exactly matches the class and constituents of the NE. This means that any NE where there is only overlap, i.e. the classified entity is shorter or longer than the annotated entity, the entity is not considered to be correct. The Precision metric correlates the total number of correct NEs as a fraction of the total number of classified NEs while Recall computes the number of correct NEs as a fraction of the total number of expected NEs, while *F*-score calculates the harmonic mean between Precision and Recall.

The NER systems for each of the languages were evaluated using 10-fold cross validation. The averaged results of these 10-fold runs are presented in Table 2. Although these results are encouraging, the results prove that there is still reasonable room for improvement. These results indicate that most of the systems perform similarly, with relatively balanced results in terms of the Precision and Recall of the systems. There are however a couple of exceptions. The SiSwati system does not perform nearly as well as the other systems, in large part due to low recall. More than half of the items not identified by the SiSwati system were classified as MISC by the annotator, some of which are foreign language strings, numbered references to laws, and publications. Although not as pronounced as with SiSwati, almost all of the other languages has some problems differentiating between the MISC class and the other available categories.

Although the two worst performing systems are both conjunctive languages, two of the other conjunctive languages perform as well or better than the disjunctive languages. Unfortunately, it is not clear why this difference exists between these languages. Even after an additional investigation of the incorrect instances, there doesn't seem to be a clear reason for the differences, and it could be a function of the annotation quality or the nature of the text in the sense that there are more loaned NEs in the isiXhosa and isiNdebele data. More details on specific error classes are provided in the following section.

Language	**Precision**	**Recall**	***F*-score**
Afrikaans	0.7859	0.7332	0.7586
isiNdebele	0.7703	0.7326	0.7510
isiXhosa	0.7860	0.7561	0.7708
isiZulu	0.7356	0.6664	0.6993
Sesotho sa Leboa	0.7612	0.7288	0.7446
Sesotho	0.7617	0.7027	0.7309
Setswana	0.8086	0.7547	0.7806
SiSwati	0.6903	0.6017	0.6429
Tshivenda	0.7396	0.7292	0.7343
Xitsonga	0.7248	0.6946	0.7093

Table 2: Evaluation results for automatic named entity recognition for South African languages

4.2 Error analysis

As the NER systems created here are not yet as accurate as one would hope, some additional analysis of the errors that occur in the output of the NER systems was performed. This will aid in identifying areas of future research that could improve the quality of these systems.

The biggest problem with the current implementation is that all of the systems miss a large number of annotated entities, with the result that none of the systems have Recall < 0.76, which negatively impacts the *F*-score of all the NER systems. Most of the Recall problems relate to general or ambiguous named entities that occur in the data, especially references to people and governmental organisations by their official designation, such as "Office of the National Prosecutor" or "Western Province Industrial Action Plan". Most of the systems also tag tokens assigned to the MISC class as OUT, thereby lowering the Recall. Many items in all of the languages are also tagged as MISC by the NER system while they should be either ORG or PERS. This also has a negative impact on the Precision of the NER systems, as most of the incorrect tagged items are marked as MISC. It may well be necessary to update the data to exclude the MISC category in order to determine whether Recall can be improved by excluding this category.

One final problematic area of NE detection in the African languages relates to the naming conventions in the culture. It is very common for names of people to be regular words in the language, rather than unique proper names, comparable to the name *Summer* or surname *Bush*. This phenomenon results in many proper names not correctly tagged by the NER systems.

5 Conclusion

This paper described part of the NCHLT Text Phase II development project, tasked with developing protocols, 15,000 tokens annotated for named entities, and automatic NER systems, for ten of the official languages of South Africa. The development of these resources provides the research and development community of South Africa with another important resource for the further development of human language technology in the South African context.

We describe the protocols and annotated data sets, as well as the automatic NER systems built from these annotated data sets. It is shown that although there are several challenges in classifying named entities for these languages, especially given the limited scope of available feature resources, relatively accurate NER systems can be constructed within the government domain.

Although these baseline systems are already usable in other, larger HLT systems, there is still significant room for improvement in various areas of the development of the NER systems. Firstly, an investigation into the impact of the MISC class on the evaluation results presented here. Additional morphological analysis will also likely be required for the conjunctive languages in order to improve the systems. Another avenue of investigation will be to construct more extensive language specific gazetteers for each of the languages. Lastly, the current NER systems should be applied to other domains to determine how well government domain NER transfers to other domains.

6 Acknowledgements

The NCHLT Text project reported on in this paper was made possible with the financial support of the National Centre for Human Language Technology, an initiative of the South African Department of Arts and Culture.

The author would also like to thank the contributors from various South African research institutions and linguistic experts that assisted on the project.

7 References

Babych, B. & Hartley, A. (2003). Improving machine translation quality with automatic named entity recognition. In Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT: Association for Computational Linguistics, pp. 1-8.

CTexT. (2015). Linguistic Annotation and Regulation Assistant - Lara3 (Version 3.0.9). Potchefstroom: North-West University.

Das, A. & Garain, U. (2014). CRF-based Named Entity Recognition@ ICON 2013. *arXiv preprint arXiv:1409.8008*.

Doddington, G.R., Mitchell, A., Przybocki, M.A., Ramshaw, L.A., Strassel, S. & Weischedel, R.M. (2004). The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In LREC, pp. 1.

Eiselen, R. & Puttkammer, M.J. (2014). Developing text resources for ten South African languages. In Proceedings of the 9th language resource and evaluation conference, Reykjavik, Iceland, pp. 3698-3703.

Fourie, W., Du Toit, J. & Snyman, D. (2014). Comparing support vector machine and multinomial naive Bayes for named entity classification of South African languages. In Pattern Recognition Association of South Africa, Cape Town, South Africa.

Grishman, R. (2003). Information extraction. *The Handbook of Computational Linguistics and Natural Language Processing*, pp. 515-530.

Grishman, R. & Sundheim, B. (1996a). Design of the MUC-6 evaluation. In Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996: Association for Computational Linguistics, pp. 413-422.

Grishman, R. & Sundheim, B. (1996b). Message Understanding Conference-6: A Brief History. In COLING, pp. 466-471.

Konkol, M. & Konopík, M. (2013). Crf-based czech named entity recognizer and consolidation of czech ner research. In Text, Speech, and Dialogue: Springer, pp. 153-160.

Kudo, T. (2005). CRF++: yet another CRF toolkit. http://crfpp.sourceforge.net.

Kushida, C.A., Nichols, D.A., Jadrnicek, R., Miller, R., Walsh, J.K. & Griffin, K. (2012). Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical care,* 50, pp. S82-S101.

Matthew, G.D. (2013). *Benoemde-entiteitherkenning vir Afrikaans.* MA Thesis, North-West University, Potchefstroom.

McCallum, A. & Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4: Association for Computational Linguistics, pp. 188-191.

Nadeau, D. & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes,* 30(1), pp. 3-26.

Puttkammer, M.J. (2006). *Outomatiese Afrikaanse tekseenheididentifisering.* MA Thesis, North-West University, Potchefstroom.

Ramshaw, L.A. & Marcus, M.P. (1999). Text chunking using transformation-based learning. In *Natural language processing using very large corpora* (pp. 157-176). Berlin: Springer.

Sundheim, B.M. (1996). Overview of results of the MUC-

6 evaluation. In Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996: Association for Computational Linguistics, pp. 423-442.

Tjong Kim Sang, E.F. & De Meulder, F. (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: Association for Computational Linguistics, pp. 142-147.

Turchi, M., Atkinson, M., Wilcox, A., Crawley, B., Bucci, S., Steinberger, R., et al. (2012). Onts: optima news translation system. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics: Association for Computational Linguistics, pp. 25-30.