South African Language Resources: Phrase Chunking

Roald Eiselen

Centre for Text Technology, North-West University, Potchefstroom Campus, South Africa

Roald.Eiselen@nwu.ac.za

Abstract

Phrase chunking remains an important natural language processing (NLP) technique for intermediate syntactic processing. This paper describes the development of protocols, annotated phrase chunking data sets and automatic phrase chunkers for ten South African languages. Various problems with adapting the existing annotation protocols of English are discussed as well as an overview of the annotated datasets. Based on the annotated sets, CRF-based phrase chunkers are created and tested with a combination of different features, including part of speech tags and character n-grams. The results of the phrase chunking evaluation show that disjunctively written languages can achieve notably better results for phrase chunking with a limited data set than conjunctive languages, but that the addition of character n-grams improve the results for conjunctive languages.

Keywords: Language resource development, South African languages, phrase chunking, shallow parsing

1 Introduction

Syntactic analysis is an important part of many natural language processing systems, but generally requires large annotated data sets that are syntactically parsed and disambiguated in order to be effective. Abney (1992) proposed an alternative method for performing partial syntactic analysis in environments where full syntactic analysis is not required, but a level of analysis beyond part of speech is required. Phrase chunking, or shallow parsing, provides a flat representation of the major syntactic categories by assigning tokens to non-recursive segments, without resolving more complex attachment (Abney, 1992; Kübler et al., 2010). This approach allows for syntactic analysis that is useful, but does not require full grammars or large training sets that are complex and expensive to develop.

This approach to syntactic analysis is especially interesting in the South African context where work on automatic syntactic analysis has been limited, with little data widely available for research or development purposes. Furthermore, at least nine of the official South African languages have limited syntactic or annotated data resources. An intermediate approach that will allow for some level of syntactic analysis and can be used other NLP technologies and applications, is ideal for this situation.

The research described in this paper gives an overview of the NCHLT Text Phase II project, which undertook the task of annotating 15,000 tokens with their phrasal constituency and creating phrase chunkers for ten of the official South African languages. This paper describes the development process for these resources by presenting an overview of some of the challenges and solutions experienced during the development.

The first part of the paper presents an overview of the annotated datasets, including the development of protocols for use during the annotation process. Some of the language specific issues and considerations encountered during the annotation process are also discussed. This is followed by a short description of the automatic phrase chunkers that were developed using conditional random fields (CRF). After these descriptions, the phrase chunkers are evaluated and we provide results showing that high quality CRF phrase chunkers for the disjunctive South African languages can be constructed with as little as 15,000 annotated tokens. The phrase chunkers for the conjunctive languages are however far inferior and will require additional data and perhaps different approaches in order to be of the same quality. Finally, considerations for future work to improve the quality of the phrase chunkers, especially for conjunctive languages, are described.

2 Background

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. As Prinsloo and De Schryver (2002) describe, the indigenous languages follow an orthography where a linguistic word is either written as multiple orthographic entities, i.e. disjunctively, or as a single orthographic entity, conjunctively.

Over the past decade the South African Department of Arts and Culture has funded a variety of projects in the domain of human language technology, with the aim of developing language resources for the official languages of South Africa. These activities aim to ensure that the indigenous languages of the country remain viable modes of communication in the digital age, as well as using language technology to make information available to users in their native language. The projects funded by the Department typically target ten of the official languages, as English is already widely researched and has a large number of language resources available.

The project to develop phrase chunking resources for the South African 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 phrase chunkers.

The project was an interinstitutional resource development project, including linguists and language experts from six different South African Universities involved in the project, while the project was coordinated by the Centre for Text Technology at the North-West University.

Although there is a large distribution of syntactic related grammar work done across the ten African languages, very little digital syntactic work is available for these languages. The work described here is one of the first efforts in creating annotated phrase chunking datasets for the South African languages and automatic phrase chunkers for these languages. All of the resources described in this work have been made available under Creative Commons Attribution Licenses via the Languages Resource Management Agency[1]; a hosting and distribution hub for NLP related language resources for the South African languages.

3 Approach

3.1 Protocols and Data Annotation

The first part of the project focused on the development of the annotated datasets. The sets are a subset of the tagged sets described in Eiselen and Puttkammer (2014), annotated for an additional level. This annotation process was facilitated through annotation protocols provided to annotators that are based on those used for the annotation of the CoNLL-2000 shared task (Tjong Kim Sang & Buchholz, 2000), localised and adapted for each language. The protocols distinguish five main types of phrases, namely Noun (NP), Verb (VP), Adjective (AdjP), Adverb (AdvP), and Prepositional (PP) phrases. Since all phrases in the phrase chunking paradigm must be non-overlapping, maximal projections, with no internal chunks, NP and VP chunks usually supersedes Adjective and Adverb phrases, while prepositional phrases consist exclusively of prepositions (Abney, 1992; Tjong Kim Sang & Buchholz, 2000).

The data annotation process also followed the well-established Inside, Outside, Beginning (IOB) labelling scheme (Ramshaw & Marcus, 1999), which can very easily be used to train an automatic labeller. Since this scheme is not ideal for annotating by hand, an extension of the Linguistic Annotation and Regulation Assistant, LARA3 (Schlemmer, 2015)[2], was developed, which assisted annotators in creating accurate annotations. The tool provides basic drag and click functionality to highlight and assign a sequence of words to a particular phrase class. The annotated data is then stored in the IOB format.

Although the annotation scheme is a well-established reference that has been widely implemented for various languages around the world, the nature of the South African languages caused several issues with regard to how specific constructions would be handled. For the various African languages there were several issues, usually due to the distinction between conjunctive and disjunctive languages.

Afrikaans, as a Germanic language with largely similar syntactic constructions to English, had relatively few problems when applying the protocol. The one exception to this is the use of double negatives in Afrikaans. In almost all cases, Afrikaans requires two negation particles when expressing negation, for example:

(1) *[NP Ek NP] [VP sal nie VP] [NP die werk NP] [VP doen nie VP].*
('I will not do the work', lit. I will not the work do not)

The original protocols called for the negation particle to be attached to the associated verb, and this could be followed in the Afrikaans data, however, in cases where the negation particle is split from the verb, only one of the particles is attached to the verb, while the other is annotated as outside.

As was mentioned earlier, disjunctive languages follow an orthography where a large number of particles and concords, that would form part of the linguistic word, are separated and form multiple orthographic words. As an example, the following phrase has only two linguistic words (separated by "/"), but four orthographic words:

(2) *Ke tla reka / nama*
(I shall buy meat)
(Example from Pretorius et al. (2009))

The consequence of this is that phrase chunking performs a task of combining these orthographic entities into groups more closely related to the conjunctive constructions used in the Nguni languages. The nominal phrase chunks followed a similar structure to that of English, with all nouns, pronouns, adjectives, and enumeratives usually included in the noun phrase. The one construct that caused a lot of discussion, but was ultimately handled in the same way as English, was the possessive construction which is very widely used in the disjunctive languages. Constructions that would typically consist of a compound noun phrase in English are typically expressed as possessives in the disjunctive languages, for example:

(3) *[NP foromo] [PP ya PP] [NP kgopelo NP]*
('application form', lit. 'form of application'

These constructions were always handled as NP-PP-NP phrase chunks.

With regards to verbal constructs, the disjunctive languages use subject and object concords and tense markers, all of which are included as part of the verb phrase as shown in (4), which consists of a subject concord, future tense marker, object concord and a verb stem:

(4) *[VP ba tlo mo swara VP]*
('They will catch him', lit. 'they will him catch'

Language	ADJP	ADVP	NP	PP	VP
Afrikaans	14	23	7350	1967	3341
isiNdebele	141	254	9575	0	2262
isiXhosa	234	493	7872	0	3468
isiZulu	605	739	7309	0	4270
Sesotho sa Leboa	0	132	5004	2611	4811
Sesotho	322	122	4817	3221	5258
Setswana	0	91	4769	2331	5062
SiSwati	270	639	7360	0	4165
Tshivenda	27	130	7849	944	2397
Xitsonga	7	52	4543	2449	5142

Table 1: Summary of annotated phrase chunking data sets

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

[2] Also available from http://rma.nwu.ac.za/

The conjunctive South African languages are highly inflectional, with one word often containing various syntactic functions, including relatives, possession, demonstratives. As an example, the isiZulu verb *baphindele* (‘go back to’) consists of three constituents, the subject concord *ba-*, verbal root *–phind-*, applicative extension *–el-* and a verb terminative *-e*.

The consequence of this is that most of the phrases in the data consist of a single word and little new information is added by providing phrasal annotations. The other major difference between the conjunctive annotated data is that because of this inflectional quality, many more adverbial phrases were annotated than is the case for either the disjunctive or Germanic languages.

Table 1 provides an overview of the distribution of phrase types for the different languages. In all, approximately 15,000 tokens were annotated for phrase chunks, and based on these phrasal chunks, automatic phrase chunkers for all of the languages were created.

3.2 Automatic Phrase Chunkers

The second part of the project focused on the development of automatic phrase chunkers for the ten languages based on the annotated data developed for the project. The nature of the IOB format lends itself well to the development of machine learning systems that treat the phrase chunking problem as a sequence labelling problem. Two techniques are typically employed to perform sequence labelling, *viz.* support vector machines and conditional random fields (Bhat & Sharma, 2011; Gune et al., 2010; Kudo & Matsumoto, 2001; Lafferty et al., 2001; Sha & Pereira, 2003). Based on the experimental results of Eiselen (2014) it was decided to train linear chain CRF phrase chunkers with L2 regularisation based on the CRF++ implementation (Kudo, 2005).

Three different chunkers for each language were trained to evaluate the impact of different features on the quality of the phrase chunkers, as detailed in Table 2. In all three cases, the phrase chunkers used a set of standard features, based on the token strings of the current token (T_n) to be tagged and the surrounding tokens. In addition to the token features, all three sets also used the part of speech tags assigned to the current token (P_n), along with the POS tokens of the surrounding context, totalling 19 features typically used in CRF-based phrase chunkers.

Because of the fact that four of the languages in the development project have a conjunctive orthography, additional experiments were performed to determine whether the use of some morphological characteristics could improve the quality of the technologies, especially for the conjunctive languages. In addition to the POS-only phrase chunker, two additional phrase chunkers were created using either the first and last N characters of the token as additional features, for N values of between one and five. For brevity only the two most successful of these are discussed in the evaluation section of this paper, namely N=2 (CN-2) and N=4 (CN-4). Each of the different systems was evaluated to determine the ideal chunker for each language, and the results of these evaluations are presented in the next section.

Feature set	Features
POS-only	T_n, T_{n-1}, T_{n-2}, T_{n+1}, T_{n+2}, T_{n-1}/T_n, T_n/T_{n+1} P_n, P_{n-1}, P_{n-2}, P_{n+1}, P_{n+2}, P_{n-2}/P_{n-1}, P_{n-1}/P_n, P_n/P_{n+1}, P_{n+1}/P_{n+2}, $P_{n-2}/P_{n-1}/P_n$, $P_{n-1}/P_n/P_{n+1}$, $P_n/P_{n+1}/P_{n+2}$
CN-2	All POS features, and: C_{start2}, C_{end2}, C_{start2}/C_{end2}
CN4	All POS features, and: C_{start4}, C_{end4}, C_{start4}/C_{end4}

Table 2: Feature sets for experimental phrase chunkers

4 Evaluation

Phrase chunkers are generally evaluated with three different evaluation metrics, namely Precision, Recall and *F*-score. The Precision metric correlates the total number of correct phrases as a fraction of the total number of assigned phrases, while Recall computes the number of correct phrases as a fraction of the total number of expected phrases. *F*-score is the harmonic mean between Precision and Recall, typically used to evaluate the overall performance of the system (Daelemans et al., 1999; Sha & Pereira, 2003).

The phrase chunkers are evaluated with 10-fold cross validation, where three systems for each language are compared, the POS systems that only include part of speech and local token context information, and the character n-gram (CN) systems that include the rudimentary character n-gram combinatory features described in the previous section. The results for the respective evaluation metrics are provided in Table 3.

Language	***F*-score**		
	POS-only	CN-2	CN-4
Afrikaans$^{\alpha}$	0.9477	0.9509	**0.9517**
isiNdebele§	0.8506	0.8726	**0.8747**
isiXhosa§	0.8340	**0.8596**	0.8545
isiZulu§	0.8813	0.9032	**0.9156**
Sesotho$^{\alpha}$	0.8455	0.8531	**0.8570**
Sesotho sa Leboa $^{\alpha}$	0.9742	**0.9777**	0.9755
Setswana $^{\alpha}$	**0.9535**	0.9497	0.9494
SiSwati§	0.7796	0.8230	**0.8380**
Tshivenda$^{\alpha}$	0.9291	**0.9366**	0.9352
Xitsonga$^{\alpha}$	0.9295	0.9241	**0.9308**

Table 3: Evaluation results for automatic phrase chunkers for South African languages.

These results indicate that there is a clear quality difference between the disjunctive ($^{\alpha}$) and conjunctive languages(§). There are two reasons for this difference. Firstly, the fact that the vocabulary of the disjunctive languages is significantly smaller (Prinsloo & De Schryver, 2002) than those of the conjunctive languages means that the conjunctive languages have far greater data sparsity in the model, thus influencing the model accuracy negatively. Secondly, one of the features of the model is the automatically assigned part of speech tag, and as Eiselen and Puttkammer (2014) reported, the quality of the conjunctive language part of speech taggers is substantially lower than that of the disjunctive languages. As a feature this is not as accurate as the feature for the disjunctive languages, possibly negatively influencing the results.

Unlike Eiselen (2014), chunkers for all but one of the languages improved across all of the metrics with the use of the additional character n-gram information, although different values of N were necessary for different languages. The *F*-score improvements for the disjunctive languages are very moderate (between 0.0013 and 0.0115), but the improvements for conjunctive languages is substantial, ranging between 0.0241 and 0.0584, which indicate that adding additional morphological characteristics for these languages is an important feature that could be further improved to produce better phrase chunkers. The reason Setswana did not improve with any of the additional features is that more than 70% of the incorrectly tagged tokens are particles, concords or markers that do not inflect and therefor do not have any morphological characteristics.

The relatively poorer performance of Sesotho stems from the incorrect tagging of verbal phrase chunks, with more than half of the token errors related to particles and concords that are either incorrectly tagged as prepositions, or there is a difference in the length of the phrase tagged by the human annotator and the phrase chunker. Additional work will need to be done to refine the training set of Sesotho to improve the quality of the performance on these verbal aspects.

5 Conclusion

This paper described a part of the NCHLT Text Phase II development project, tasked with developing protocols, 15,000 phrase chunk annotated tokens, and automatic phrase chunkers 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. Even as baseline systems with limited annotated data, the chunkers for the disjunctive languages and Afrikaans perform well with all but one attaining *F*-scores of above 0.93. The chunkers for the conjunctive languages will require additional work to improve their performance, but the use of character n-gram data is shown to make significant improvements to the systems, and more complex morphological characteristics may further these improvements.

The bulk of future work for these phrase chunkers is an investigation of the impact of different morphological features, including lemmas, stems, roots, morphological decompositions and full morphological analysis. Secondly, the quality of the chunkers should be validated on a broader test set, since the current sets are limited to the government domain and the models may not scale as well to different 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 Bibliographical References

Abney, S.P. (1992). Parsing by chunks. In R. C. Berwick, S. P. Abney & C. Tenny (Eds.), *Principle-based parsing* (pp. 257-278). Berlin: Springer.

Bhat, R.A. & Sharma, D.M. (2011). A hybrid approach to kashmiri shallow parsing. In The 5th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC-2011).

Daelemans, W., Buchholz, S. & Veenstra, J. (1999). Memory-based shallow parsing. *arXiv preprint cs/9906005*.

Eiselen, R. (2014). Phrase chunking for South African languages: an investigation for Sesotho sa Leboa, Setswana, and Afrikaans. In Pattern Recognition Association of South Africa, Cape Town, South Africa.

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.

Gune, H., Bapat, M., Khapra, M.M. & Bhattacharyya, P. (2010). Verbs are where all the action lies: experiences of shallow parsing of a morphologically rich language. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters: Association for Computational Linguistics, pp. 347-355.

Kübler, S., Beck, K., Hinrichs, E. & Telljohann, H. (2010). Chunking German: an unsolved problem. In Proceedings of the 4th linguistic annotation workshop, Uppsala, Sweden: Association for computational linguistics, pp. 147-151.

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

Kudo, T. & Matsumoto, Y. (2001). Chunking with support vector machines. In Proceedings of NAACL-HLT 2001, Pittsburgh, PA: Association for computational linguistics, pp. 1-8.

Lafferty, J., McCallum, A. & Pereira, F.C. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th international conference on machine learning, Williamstown, MA, pp. 282-289.

Pretorius, R., Berg, A., Pretorius, L. & Viljoen, B. (2009). Setswana tokenisation and computational verb morphology: Facing the challenge of a disjunctive orthography. In Proceedings of the First Workshop on Language Technologies for African Languages: Association for Computational Linguistics, pp. 66-73.

Prinsloo, D. & De Schryver, G.-M. (2002). Towards an 11 x 11 array for the degree of conjunctivism/disjunctivism of the South African languages. *Nordic Journal of African Studies,* 11(2), pp. 249-265.

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.

Schlemmer, M. (2015). Linguistic Annotation and Regulation Assistant - Lara3 (Version 3.0.9). Potchefstroom: North-West University - Centre for Text Technology.

Sha, F. & Pereira, F. (2003). Shallow parsing with conditional random fields. In Proceedings of NAACL-HLT 2003, Edmonton, Canada: Association for Computational Linguistics, pp. 134-141.

Tjong Kim Sang, E.F. & Buchholz, S. (2000). Introduction to the CoNLL-2000 shared task: chunking. In Proceedings of the 4th Conference on Computational Natural Language Learning, Lisbon, Portugal: Association for Computational Linguistics, pp. 127-132.