Processing Parallel Text Corpora for Three South African Language Pairs in the Autshumato Project

Hendrik J. Groenewald, Liza du Plooy
Centre for Text Technology (CTexT™),
North-West University (NWU), Potchefstroom Campus, South Africa
{handre.groenewald, liza.duplooy}@nwu.ac.za

Abstract

Multilingual information access is stipulated in the South African constitution. In practise, this is hampered by a lack of resources and capacity to perform the large volumes of translation work required to realise multilingual information access. One of the aims of the Autshumato project is to develop machine translation systems for three South African languages pairs. These machine translation systems are envisaged to serve as tools that will enable the human translator to do his/her work more efficiently and thereby indirectly contributing to improving access to multilingual information services. This paper describes the gathering and processing of parallel corpora for the purpose of developing machine translation systems. We highlight the difficulties in gathering parallel corpora and provide detailed information about the anonymisation of parallel corpora. We also describe various data processing methods to prepare the parallel corpora for use in the development of machine translation systems.

1. Introduction

South Africa is known for diversity in cultures, languages and religious beliefs. The South African constitution (Republic of South Africa, 1996) recognises no fewer than eleven official languages. The South African government aims to meet the constitutional obligations on multilingualism, by providing equitable access to government services, knowledge and information (Republic of South Africa, 2003). Promoting access to information in all of the official South African languages is a noble idea, but it is difficult to apply in practise, due to large volumes of translation work that are required to realise this.

It is therefore no surprise that government translation agencies such as the National Language Service (NLS) are struggling to keep up with large volumes of translation work. This problem is aggravated by the fact that machine-aided translation tools and resources are often not available for South African languages. Another contributing factor is the fact that proprietary computer-assisted translation software suites are not widely used by government translation agencies due to high licensing costs and incompatibilities with open source computer operating systems.

The consequence of this is that English often acts as the *lingua franca* (although it is the mother tongue of only 8.2% of the South African population (Van der Merwe and Van der Merwe, 2006)), with the unfortunate result that the remaining indigenous South African languages are yet even further marginalised. The dominance of English has the resulting effect that a large number of South African citizens are deprived of their constitutional right of access to information in their language of choice. Innovative solutions are required to ensure multilingual access to information in South Africa. The South African Government is aware of the potential impact that Human Language Technology may have in this regard and is therefore involved in a number of initiatives and projects to promote multilingualism through the use of Human Language Technology. One such a project is the Autshumato Project, which is discussed in Section 2. The rest of this paper is organised as follows: In Section 3 we provide more information about the machine translation approach that is followed in the Autshumato project. Section 4 focuses on data providers, while Section 5 provides detailed information about the anonymisation system that we have developed. Further processing of parallel corpora is discussed in Section 6. The paper concludes in Section 7 with some directions for future work.

2. Autshumato Project

The Autshumato[1] project was commissioned in 2006 by the South African Department of Arts and Culture. The aim of the project is to develop open source machine-aided translation tools and resources for South African languages. One of the interesting outcomes of the Autshumato Project is the development of machine translation systems for three South African languages pairs, namely English – > isiZulu, English – > Afrikaans, and English – > Sesotho sa Leboa. The purpose of these machine translation systems is to provide government translators with a tool that may help them to perform their work more efficiently and increase consistency and productivity. Obtaining and processing parallel data to develop machine translation systems for three South African Language pairs is the central theme of this paper.

3. Machine Translation

Interest in machine translation as an area of research started to gain momentum after World War II. The complexity of machine translation was grossly

[1] Autshumato was a Khoi-khoi leader that worked as in interpreter between Europeans and the Khoi-khoi people during the establishment of the Dutch settlement at the Cape of Good Hope in the 17th century (Giliomee and Mbenga, 2007).

underestimated in those early years and together with the publication of the ALPAC report (ALPAC, 1966) resulted that interest in (and funding of) machine translation research gradually decreased. Despite the pessimism of the ALPAC report, machine translation research continued and a large number of machine translation systems have been developed over the years with varying degrees of success (De Pauw *et al.*, 2009).

The rapid expansion of the internet and computer processing power in the 1990's, together with the resulting increase in the availability of electronic data led to an increased interest in the use of the statistical machine translation (SMT) method. The performance of statistical machine translation systems is to a large extent dependent on the amount of and quality of parallel text corpora available.

All of the official South African languages, except English, are considered to be resource scarce languages. The availability of parallel text corpora for these languages is limited. The lack of parallel text corpora implies that alternative methods to the mere addition of parallel data must be sought to improve the quality of the machine translation systems that are being developed in the Autshumato project. The approach followed in the Autshumato project is to utilise statistical machine translation to create baseline machine translation systems that can be augmented with rules based on expert linguistic knowledge. It was decided that SMT would form the basis for the development of machine translation systems in the Autshumato Project for the following reasons:

1) SMT is currently the preferred approach of numerous industrial and academic research laboratories.
2) State-of-the art open source SMT toolkits are readily available.
3) Less expert linguistic knowledge is required to create a working baseline system in comparison to rule-based systems.

4. Data Providers

As mentioned in the previous section, it is difficult to obtain parallel text corpora for the language pairs for which machine translation systems are currently being developed in the Autshumato project. Apart from government sources, few other sources of parallel text corpora exist. It is often necessary to approach private translation companies and freelance translators to obtain parallel text corpora. The unavailability of parallel text data can be ascribed to the following reasons:

1) Computer-assisted translation software suites are not widely used, with the result that translation memories are not readily available for the indigenous South African languages.
2) Lack of publications like books, newspapers, magazines and websites in the indigenous South African Languages.
3) Lack of sound document management practices, which makes it difficult to obtain parallel documents from translators.
4) Unwillingness of translators and private companies to make their data available for purposes of machine translation research.

5. Text Anonymisation

5.1 Introduction

As mentioned in the previous section, several potential data providers are unwilling to provide their parallel corpora for the development of machine translation systems. One of the main reasons for this is that their corpora often contain confidential information. For this reason, we are developing text anonymisation software to identify confidential information in parallel corpora. We will then anonymise the corpora by replacing the entities conveying the confidential information, by randomly selected entities from the same category.

Text anonymisation can be seen as a subcategory of named entity recognition. It differs from named entity recognition in the respect that it focuses only on the entities in text containing information of a confidential nature.

Neamatullah *et al.* (2008) did research on the anonymisation of medical records and developed a *Perl*-based de-identification software package that can be used on most free text medical records like notes from nurses. It was developed using gazetteers (i.e. dictionaries containing specialised entity lists), regular expressions and simple rules. This software was exclusively developed for use in the medical domain.

Text anonymisation software is being developed as part of this research for English, Sesotho sa Leboa, isiZulu and Afrikaans (the four languages involved in the development of the Autshumato machine translation systems). Research on named-entity recognition for Afrikaans was done by Puttkammer (2006) as part of the development of an Afrikaans tokeniser. A named-entity recogniser for isiZulu was developed as a part of the project *REFLEX: Named Entity Recognition and Transliteration for 50 Languages* (Sproat, 2005). For Sesotho sa Leboa research has been done on the recognition of spoken personal names as a part of the *Interactive Voice Recognition (IVR)*-system (Modipa, Oosthuizen & Manamela, 2007).

5.2 Method and Implementation

A rule-based approach was followed in the development of the anonymisation system, using gazetteers, regular expressions and simple context rules. A proof-of-concept system was first developed for Setswana, Afrikaans and English. This system has been recently adapted to include isiZulu and Sesotho sa Leboa. We aimed to make the anonymisation system as language independent as possible, to ensure rapid expansion to the remaining official South African languages in the future. The operation of the anonymisation system is as follows: Firstly, entities with a predictable form like dates,

addresses, contact numbers, email addresses and URL's are marked with regular expressions. Next, all the words that appear in any of the gazetteers are marked. Lastly, context rules are applied to find entities that do not appear in any of the gazetteers. A logic flow of the various steps in the anonymisation process is shown in Figure 1. Each of these steps will now be discussed in more detail in the rest of Section 5.2.

5.2.1 Regular Expressions

As mentioned in the previous section, regular expressions are used to identify entities with a predictable form. Dates are recognised by combining regular expressions with lists containing words that are regularly found in dates, like the names of months and days, in the different languages (see step 3 in Figure 1). Dates in formats like *2009-09-27, 05/12/2010* and *16 Feberware 1978* are also identified. To recognise addresses, lists with location names are combined with regular expressions. Due to the numerous valid formats of addresses, several regular expressions are implemented to recognise the different formats and types of addresses. Contact numbers are written differently in different parts of the world, but for the goal of this research specific attention has been given to numbers in the formats commonly used in South Africa.

5.2.2 Gazetteers

The gazetteers were assembled from a variety of sources and are continuously updated. All of the words in the text that appear in any of the gazetteers are marked by the system (see step 6 in Figure 1).

First Names and Surnames

The list of first names of Puttkammer (2005) was expanded by the addition of names that are found amongst to various ethnic groups in South Africa. The list currently consists of 8,853 unique names. The list of surnames used by Puttkammer (2005) consists of 3,174 surnames. The list used by Neamatullah et al. (2008) is freely available on the web and was added to the existing list of Puttkammer. Surnames were also extracted from the address book of the North-West University, available on the university's website (North-West University, 2004). Several of the first names and surnames contained in the above-mentioned lists are also valid words in the indigenous languages when not used in the "first name or surname sense". For example, the Setswana sentence *Ke na le khumo* means "I have wealth", but *Khumo* can also be a first name. These words were removed from the list by comparing it to lexica consisting of valid lower case words of the different languages. The final list consists of 81,711 surnames.

Company, organisation and product names

Once again the list used by Puttkammer (2006) was expanded to be used in the anonymisation system. A small amount of company names were obtained from the website of the Johannes Stock Exchange (JSE, 2005) and appended to the list of Puttkammer (2006). This list needs to be expanded considerably in future.

5.2.3 Context Rules

Context rules are applied to find entities that do not appear in any of the gazetteers (step 9 in Figure 1). An example of a context rule is: "if a word following a word that has been tagged as a first name starts with a capital letter, and if that word does not appear in one of the lexica, it is considered to be a personal name".

The last step is to compare every capital letter word against lowercase lexica. If the particular word does not appear in any of the lexica, it is considered to be a named entity (see step 11 in Figure 1). This has a positive effect on the recall of the system, but it also has a negative effect on precision. For the purposes of our research, high recall is more important than high precision, since we do not want any named entities that can convey confidential information to go unnoticed by the system.

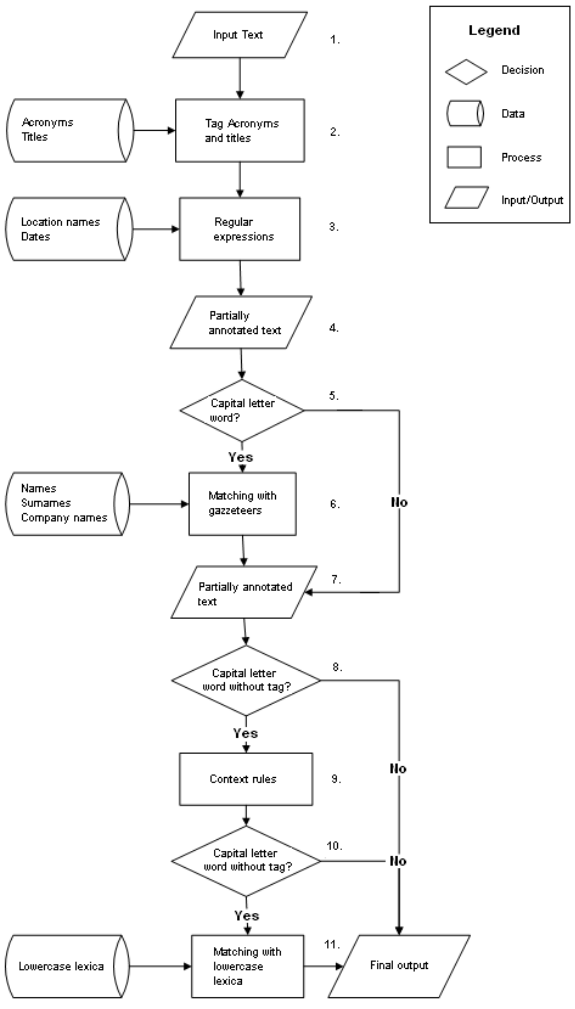

Figure 1: Logical flow of the anonymisation system

5.3 Results

The anonymisation system was tested on data from the government, university and agricultural domains. The test data were annotated manually. The results obtained for Setswana, Afrikaans and English are shown in Figure 2. Setswana obtains remarkable lower scores than Afrikaans

and English. One reason for this is the fact that many first names and surnames in Setswana (and many other indigenous South African and African languages) are also valid words when not used in the "first name or surname" sense of the word (see Section 5.2.2).

	English	Afrikaans	Setswana
Recall	0,796	0,799	0,777
Precision	0,832	0,805	0,756
F-Score	0,814	0,802	0,767

Table 1: Results obtained by the anonymisation system

6. Processing of Data

6.1 Sentencisation

After the parallel text corpora is obtained from the various data suppliers, the data must be sentencised (i.e. split into sentences) before it can be aligned at sentence level. Sentencisation algorithms for all the official South African languages have been developed as part of the Autshumato project. These sentencisation algorithms are based on a combination of language specific rules and abbreviation lists.

6.2 Alignment

Once the data have been sentencised, the alignment process can begin. Microsoft's bilingual sentence aligner (Moore, 2002) is used to automatically align sentences. The sentences that are not aligned during the automatic alignment process are manually aligned. After the parallel corpora have been aligned at sentence level, it is ready to be used to train statistical machine translation systems. The current number of aligned text units is shown in Table 2.

Language Pair	Aligned Text Units
English-Afrikaans	439,890
English-Sesotho sa Leboa	50,238
English-isiZulu	92,560

Table 2: Number of aligned text units

7. Conclusion and Future Work

In this paper we motivated the importance of developing machine translation systems for South African languages. We described the problematic aspects associated with the gathering of parallel corpora for resource scarce South African languages. We also focused extensively on the anonymisation algorithm that has been developed to motivate data suppliers to make their data available to the Autshumato project. We concluded with a brief description of the processing steps that are carried out on the data after the anonymisation process.

Future work includes the improvement of the performance of the current anonymisation system. Possible ways to achieve this is to expand the gazetteers (especially the company and product names), "cleaning" the gazetteers by removing ambiguous words, adding more context rules and refining the existing rules. We also want to augment the current system with machine learning techniques to determine if further improvements in recall and precision can be obtained.

8. References

ALPAC. (1966). *Languages and machines: Computers in translation and linguistics.* A report by the Automatic Language Processing Advisory Committee, Division of Behavioural Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences, National Research Council. Publication 1416.

Bekker, A., Groenewald, H.J. & McKellar, C.A. (2009). *Progress Report: DAC Project B: Machine Translation Systems*. Potchefstroom: CTexT.

De Pauw, G., Wagacha, P.W. & de Schryver, G. (2009). The SAWA Corpus: a Parallel Corpus English-Swahili. In *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages*. Athens: Association for Computational Linguistics, pp 9–16.

Giliomee, H. & Mbenga, B. (2007). *New History of South Africa*. Cape Town: Tafelberg.

Johannesburg Stock Exchange. (2009). Listed Companies. [Online]. Available: http://www.jse.co.za/listed_companies.jsp (accessed October 2009).

Modipa, T., Oosthuizen, H. & Manamela, M. (2007). Automatic Speech Recognition of Spoken Proper Names. Sovenga: University of Limpopo.

Moore, R.C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas*, volume 2499 of Lecture Notes in Computer Science, Berlin: Springer Verlag, pp. 135–144.

Neamatullah I., Douglass M.M., Lehman L.W., Reisner A., Villarroel M., Long W.J., Szolovits P., Moody G.B., Mark R.G. & Clifford G.D. (2008). Automated de-identification of free-text medical records. In *BMC Medical Informatics and Decision Making* 2008, 8(32).

North-West University. (2009). [Online]. Available: http://galactrix.puk.ac.za/opendir.asp (accessed October 2009).

Puttkammer, M.J. (2006). *Outomatiese Afrikaanse Tekseenheididentifisering "Automatic Text Unit Identification"*. Potchefstroom: NWU.

Republic of South Africa. (1996). *Constitution of the Republic of South Africa*. Act 108 of 1996. Pretoria: Government Printer.

Republic of South Africa. (2003). *National Language Policy Framework*. Pretoria: Department of Arts and Culture.

Sproat, R. 2005. REFLEX: Named Entity Recognition and Transliteration for 50 Languages. [Online]. Available: http://serrano.ai.uiuc.edu/reflex (accessed December 2009).

Van der Merwer, I.J. & Van der Merwe, J.H. (2006). *Linguistic Atlas of South Africa.* Stellenbosch: Sun Press.