University of Vlora
Technical Report

On the current state of Albanet and related applications[1]

1 Introduction

This document describes the current state of the Albanet [4] project. Albanet is a lexical database for the Albanian language, based on the Princeton Wordnet [2] database model. So far the following tasks have been completed:

1. 4600 base concepts of the Albanian language have been developed into XML synsets. This database can be downloaded from http://albanet.univlora.edu.al. A data browsing interface is also available.
2. A *stemmer* and a *contextual spell checker* for various NLP tasks in Albanian has been developed.

Much remains to be done. The rest of this report will describe the motivations, objectives and the techniques we will be using to achieve our goals.

2 Current state of Albanet

Albanet is a database of 4600 XML synsets modeled after the Princeton Wordnet [2] format and preserving the same foreign keys linking the Albanian synsets to their English correspondents.

WordNet [2], is a project that started in 1985 under the direction of professor George A. Miller at Cognitive Science Laboratory of Princeton University. This project has thus far received over $3 million in funding, mainly from government agencies interested in machine translation. However the uses of Wordnet extend beyond machine translation, from the areas of *information retrieval*, *automatic text classification* and on to such odd applications as automated crossword puzzle generators and solvers. In recent years this project has been led by Dr. Christiane Fellbaum.

Albanet has been built from scratch following closely the Princeton Wordnet model, by linking the 4600 English base synsets to their equivalents in Albanian. Currently, this first version of Albanet is not suitable for any practical application, however it is good enough to give an idea of what can be done. A complete version of Albanet is our main long term goal.

3 Using Albanet for machine translation

We have applied this limited resource to develop a number of applications as *proof of concept*, the principal application being a translation engine from Albanian to English. The method we used for machine translation tasks from Albanian to English, is generally based on the *corpus based translation* techniques (also known as *example based*) which fall under the general heading of Statistical machine translation (SMT) [5].

As of 2008, this approach to machine translation is by far the most widely-studied machine translation paradigm.

The ideas behind statistical machine translation come out of information theory. Essentially, the document is translated on the probability $p(e|f)$ that a string e in native language (for example, English) is the translation of a string f in foreign language (for example, Albanian).

In most cases these probabilities are estimated using techniques of parameter estimation.

The Bayes Theorem is applied to $p(e|f)$, the probability that the foreign string produces the native string to get $p(e|f) \propto p(f|e)p(e)$, where the translation model $p(f|e)$ is the probability that the native string is the translation of the foreign string, and the language model $p(e)$ is the probability of seeing that native string. Mathematically speaking, finding the best translation $\tilde{e}$ is done by picking up the one that gives the highest probability:

$\tilde{e} = arg\max_{e \in e^*} p(e|f) = arg\max_{e \in e^*} p(f|e)p(e)$.

The success our technique also relies on the existence of a grammatically correct and complete corpus of the target language in question, composed of words, phrases and sentences from the said language.

The corpus is processed by us into a graph $G(v, e)$, which has the words of the corpus as vertices (v) and word relationships in terms of proximity in a sentence as edges (e). See *Figure 1* for an example graph representing the preceding sentence.

This graph is stored in a MySQL database using three tables, one for the vertices (words), one for the original sentences and the third table for the graph edges.

The place where *Albanet* will come in handy in this process, it at the point of using the synonymical dictionary to increase the match rate against corpus data.

Given a sentence in the source language (eg. English) we will break it down into words and each word will be mapped using Albanet as a one-to-many match (the word will be mapped to all its possible synonyms in the target language).

[1] Albanet is the acronym we use for the Albanian Wordnet project. This project falls under the general heading of *Computational Linguistics* which is the science of processing human language using computer machinery.

Then we will permute these group words to produce sentences in the target language.
Therefore for every input sentence we will use Albanet to produce a number of candidate sentences in the target language which would then in turn will be matched against the corpus graph G for the best match.
Generally we maintain that this is the right approach because it is language independent (the same algorithms can be applied with minor modifications to translation engines for other languages) and there is sufficient corpus material provided by major Albanian language translation companies [1] to make this investigation worthwhile. In the future we can extend this algorithm on other languages that are part of the interlingual index [3].
The computational complexity of our graph based algorithms remains a challenge we wish to solve in the best possible way.
Databases like *Albanet* have come in handy at improving statistical translation engines and have been generally supported by government agencies interested in this type of technology. There are a number of other tools that can be derived from *Albanet* to further improve our results.
For example, some word-sense disambiguation techniques could be applied later on to further improve our algorithm; currently we have not applied these techniques yet the the proof of concept applications.
Naturally we must first solve the problem of providing a complete version of *Albanet*

3.1 Stemming

Stemming, especially for inflective languages such as Albanian is an important component for a Natural Language Processing system. It involves finding the *root* of any given word. Our stemmer is a database driven one.
Our database of stem words and their inflective forms contains over 500,000 words, which gives our stemmer a pretty good coverage.

3.2 Sentence Similarity Algorithm

The crux of our translation engine depends on the ability to pick the best candidate sentence that most approximates the structure and meaning of the input sentence.
For example: Let's say we wish to translate the following English sentence into Albanian:
beautiful day
a simple word by word translation using *Albanet* as a dictionary would read:
i bukur dite
which would sound strange to a native speaker of Albanian.
Next we will search in our corpus for the sentence best approximating this word by word translation to reach at the correct phrase which is
dite e bukur.
There are other challenges to this technique however, namely when it comes to matching sentences containing particular words such as names of people.
Our graph techniques will come in handy here to point out such words, and substitute them into the sentence structure where appropriate.

A simple example will better illustrate this point. Suppose we have this input sentence:
John likes apples
a word by word translation will produce:
John pelqej molla
after the words *pelqej* and *molla* have been stemmed, the algorithm will pick up the fact that *John* is a singular noun.
If we have a sentence or phrase in the corpus that reads:
Timi pelqen mollet
then the singular noun in the corpus sentence will be replaced by the singular noun in the input sentence and the resulting output will be correct.
Naturally we will need to take into account the inflective forms of singular nouns as well.

4 Using Albanet synsets to extend the sentence similarity algorithm

The power of Wordnet databases relies on their exact definitions of synonymy and meronymy. We can use the first feature to extend our algorithm by matching not just against a given input word, but also by permuting its closest synonyms into the sentence.
We can also use meronymy to derive the sentence best matching a given word's context, by analyzing the corpus graph and the preceding and consecuitive words in the input sentence.

Synonymy and meronymy can both be used in the sentence similarity algorithm for producing a better match, when a sentence containing a synonym instead of the actual input word is a more likely path in the corpus Graph.

5 Dealing with spelling errors

While the corpus is considered to be grammatically correct from the get-go, such a assumption can not be made about the input. We have therefore implemented a contextual spell checker for the Albanian language.

The spell checker works by first indentifying a word that is not in the corpus, hence it is most likely a spelling error, and then produces a list of possible matches.

Then it looks at the corpus graph, to find which is the most likely word by comparing neighboring words to the possible spelling error, with likely correct sentence paths in the graph.

6 Implementation Notes

Our goal is to implement all these algorithms as a web service. One of the major challenges remains scalability, increasing the corpus size increases the efficacy of our method but at the same time increases the computational complexity.
Given the limitations of our computer hardware have implemented a *proof of concept* application using a limited corpus derived from a specific language domain.
All our work is based on open source software and tools, and in the future we wish to co-opt community's help in improving and extending our work.

7 Final notes

The implementation of these tasks requires careful investigation of various NLP techniques to decide on the best solution. Our group has held various discussions on our approach and further contact has been made with various researchers in this field from the University of Bari, Princeton University and University of Ottawa.
We wish to especially thank Dr. Christine Fellbaum for her valuable advice and guidance in this process. In the future we hope to be able to extend our work into a complete and valuable resource for the Albanian language that could give rise to a number of practical applications such as the ones described here.
Currently our proof of concept applications are available as online tools:

1. Meaning-To-Word Dictionary for Albanian: http://fjalor.kerkoje.com
2. Translation Engine Albanian-English: http://perkthe.kerkoje.com
3. Semantic information management/retrieval system: http://vlora.kerkoje.com

This report was prepared by Ervin Ruci : eruci@univlora.edu.al

Appendix A

Journal reference grid

Journal name	Website
The Association for Computational Linguistics	http://www.aclweb.org/
Language Technology World	http://www.lt-world.org/
Computational Linguistics	http://www.mitpressjournals.org/toc/coli/34/3
Statistical Machine Translation	http://www.statmt.org/

Appendix B

Further reading

1. Machine Translation of Languages, MIT Press, Cambridge, MA. - W. Weaver (1955). Translation (1949).
2. The mathematics of statistical machine translation: parameter estimation. Computational Linguistics, 19(2), 263-311. - P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer (1993).
3. Statistical phrase based translation. In Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL). - P. Koehn, F.J. Och, and D. Marcu (2003).
4. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05). - D. Chiang (2005).
5. A Model of Competence for Corpus-Based Machine Translation - Michael Carl - 2002
6. A statistical approach to machine translation Brown, Cocke, et al. - 1990
7. Toward memory-based translation Sato, Nagao - 1990
8. Automated dictionary extraction for knowledge-free examplebased translation Brown - 1997
9. Inducing Translation Templates for Example-Based Machine Translation Carl - 1999
10. Example-based machine translation using connectionist matching McLean - 1992
11. Automated Dictionary Extraction for quot;Knowledge-Freequot; Example-Based Translation Brown - 1997
12. What is a Theory of Meaning Dummett - 1975
13. Machine Translation, How Far Can It Go Nagao - 1989
14. Example-Based Machine Translation: An Adaptation-Guided Retrieval Approach Collins - 1998
15. Integrating machine translation into translation memory systems Heyn - 1996
16. A Research Perspective on How to Democratize Machine Translation and Translation Aides Aiming at High Quality Final Output Boitet - 1999
17. A statistical approach to machine translation Cockc, Vincent - 1990
18. Integrating machine translation into translation memory systems Hcyn - 1996
19. Machine Translation ftbw Far Can It Go Nagao - 1989
20. http://en.wikipedia.org/wiki/Statistical$_{m}achine_{t}ranslation$

References

[1] Albaglobal. Albaglobal traslations. http://albaglobal.com.

[2] C. Fellbaum (ed). Wordnet: An electronic lexical database. http://wordnet.princeton.edu, 1998.

[3] P. Vossen (ed.). Eurowordnet: A multilingual database with lexical semantic net- works. Kluwer Academic Publishers, Dordrecht, 1998.

[4] E. Ruci (et al). Albanet: A lexical database for the albanian language. http://fjalnet.com.

[5] W. Weaver. Machine translation of languages, 1955.