RESOLUTION OF GEOGRAPHICAL NAMES IN INFORMATION EXTRACTION

AMAL ALSHAHRANI

University of Manchester, School of Computer Science
amal.alshahrani@gmail.com

Abstract: This study is concerned with using a machine learning method for extracting geographical named entities from the unstructured text and using a GeoNames database to differentiate between place and non-place references from the recognized geospatial named entities. It also resolves ambiguous place names (e.g. a place name referring to many places) into unambiguous place references using contextual clues and other assumptions. The GENIA tagger is used to annotate the text, and the YamCha is used to train and test the data sets. To evaluate our results, we have also used the CAFETIERE system on the test data sets. The evaluation process shows that our results have a high recall and has an approximately equivalent F-measure with the evaluation results confirming the validity of the chosen method.

Keywords: GeoNames, GENIA tagger, YamCha, CAFETIERE system.

I. INTRODUCTION

Nowadays, the internet is a major source of knowledge for humans. The amount of data available on the Internet is exponentially increasing day by day. The user can easily get data from any website on the Internet. For example, he can obtain data from Wikipedia as web pages, and he can get another data from scientific websites (e.g. Geojournal) as electronic papers. There is a need for a technology or a tool to analyse the data obtained from the Internet, get information, and acquire knowledge from this data. This information and knowledge can be later used for different purposes.

Information extraction[1][2] is the process of searching a text for information. There are two main approaches for designing an information extraction system, *knowledge engineering*, and *automatic training*. Each of these two approaches has its own advantages and disadvantages and is used based on the resources available to a designer of the system.In this research, we are going to follow the *automatic training* approach to address the problem of extraction and resolution of geographical names from the unstructured English text. The geospatial data is very important nowadays for many purposes.

For example, GeoNames website is gaining popularity for helping people to search andlocate geographical entities over the web. However, some places occur as common words, or they exist repeatedly across the globe. If someone typed in, for instance, the word "Research" in http://www.geoNames.org, he/she'll find Research Australia, Victoria populated place S 37° 42' 0" E 145° 11' 0"). He/she also will find that many place names occur repeatedly across the globe. For many applications, it is important to know (1) whether a string is a place name reference at all, and (2) if it is the case, then which specific place it represents?

1. Background

Two approaches are used to extract information from the text: *knowledge engineering approach*, an *automatic training approach*. These approaches are described in the following sections.

Knowledge Engineering Approach[3]: In the knowledge engineering approach, the information extraction is performed by manually constructing and writing a set of rules. This approach is also called the rule-based approach. Typically, information extraction rules are developed by a domain expert, called a "knowledge engineer". On the other hand, the designer should be familiar with the formalism of the target system to be used. Usually, the knowledge engineer identifies texts of interest from the selected domain and then the designer identifies common patterns by using her or his intuition to develop the corresponding rules. The rules are then implemented in the information extraction system which interprets and uses them to extract useful facts from the given text. The rule-basedapproach for extracting information requires time and effort to manually develop the rules and check them against a number of texts in order to verify the correct development of rules for the desired results.

Automatic Training Approach[3]: The automatic training approach for information extraction, also called the "machine learning approach", does not involve the manual design and development of the rules. Instead, this approach uses statistical methods which automatically extract rules from the texts given as the training data. However, a large amount of training data is needed to accurately identify the rules. In addition, manual annotation of the training data is required before executing the algorithm. This typeof learning is also called "supervised machine learning".The automatic

training approach includes algorithms such as decision trees, Hidden Markov Models (HMM) [4], Condition Random Fields (CRF) [5][6][7], and Support Vector Machines (SVM) [8].

II. DETAILS EXPERIMENTAL

3.1 Methodology

The methodology comprises four stages:

- Collecting data and preprocessing:In this stage, unnecessary data from each article will be manually removed. We have built a corpus consisting of articles taken from Geojournal[1] and Wikipedia[2].
- Text annotations (Annotator): The objective of the Pre-processing stage is to clean the articles which will make our corpus. This pre-processing is done manually for each article. The objective of the Annotator is to split the input article into word, POS, Chunk and NEs tags. To achieve this object, the GENIA tagger[3] has been used.
- GeoExtractor component: The objective of the GeoExtractor is to use the output of the Annotator to generate the candidate GeoNames. YamCha tool[4] has been used to achieve this objective. Then, a GeoNames database (e.g. CAFETIER database) is fed from the output of the YamCha to confirm the GeoNames obtained by YamCha[9].
- GeoResolver component: The goal of the GeoResolver is to resolve the geographical ambiguity of the GeoName extracted from the GeoExtractor (i.e. To differentiate between ambiguous and non-ambiguous GeoNames). One way to accomplish this is to use the context surrounding each GeoNames and other assumptions.

This study focused on extracting geospatial NEs from the unstructured text. To train our data, we have built a corpus consisting of articles taken from Geojournal and Wikipedia. We have chosen Wikipedia and Geojournal because they are easy to access and to use. As they contain many types of articles, we only focused on the ones containing lots of geospatial location names (e.g. states, and cities' names). From Wikipedia, we selected 30 articles written about wars and battles because they are full of geospatial names. From Geojournal, we have selected 30 articles about GeoHealth (EthnoBiology and EthnoMedicine) which contain many of GeoNames.Prior to the Annotation process, since we only focus on the text of each article, we need to remove the specific mark-up signs, and other information that does not relate to the aim of our project.

[1]http://www.springer.com/social+sciences/population+studies/journal/10708

[2] http://en.wikipedia.org/wiki/Wikipedia

[3] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

[4] http://chasen.org/~taku/software/yamcha/

3.2 Pre-process of the corpus

Preprocessing steps of Geojournal articles:

To prepare the Geojoural article the following steps should have followed: Removing header-note and footnote, Removing an article title, Removing authors' names and their affiliations, Removing figures, images and tables, and Removing references.

Preprocessing steps of Wikipedia articles:

To prepare the Wikipedia articlethe following steps should have followed: Removing the end of each article, Removing HTML markup from the page, Removing info boxes in the Wikipedia pages, Removing Wikipedia-format links, and Removing special symbols.

3.3 Guidelines of Annotation:

To annotate our data set, we have followed a number of guidelines[5]. These guidelines are described below.

1. **Continent**:Any continent name (Asia, South America, Australia, Europe, North America, Africa, and Antarctica)
2. **Nation**:Any nation entity (e.g. UK, USA, KSA, U.S and Soviet Union... etc.)
3. **State-or-Province**: A word representing a state and a province is tagged by LOC.
4. **County-or-District**: Any word representing a county or a district is tagged.
5. **Cluster**: Will annotate the entities representing a group of countries. Examples of these entities include Eastern Europe, the European Union, the Middle East, Southeast Asia, and Latin America.
6. **Nested Region Names**: Nested region names will be tagged as one entity (i.e. location) per each region.
7. **Ref.Location**:Annotate the entities that refer to a geographic position.We annotate to the location from this type as illustrated by the following example: where the parent is always a country, and the child is always a city, village, or a province in this country.
8. **Land-Region-natural**: We annotate the entities that related to the non-artificial locations such as Caucasus.
9. **Region-International**: The entities that cross the borders of the national have been tagged.
10. **River, valley, and lake**: Any river, valley, or lake name is tagged.

There are some vague, which should be avoided while selecting the Annotation method. This is explained below[6].

Military words:such as, *Iraqi troops*

People words: such as, *Chinese*people

Government words: such as,*Russian*government

[5] http://projects.ldc.upenn.edu/ace/annotation/

[6]http://projects.ldc.upenn.edu/ace/docs/English-Entities-Guidelines_v6.6.pdf

Organization words: such as, *U.S.*Fish and Wildlife Service.

3.4 Annotation process

We have used GENIA tagger to annotate the data, The GENIA tagger is not designed for annotatingGeoNames. GENIA tagger is a tool which takes English text as an input and produces a table consist of four columns*: word*, *POS* (*Part-Of-Speech*)[7], *Chunk* and *NE*s tag. GENIA annotates the data by splitting the sentence into words.Each word then is tagged either "O: out of named entities", or other symbols related to biomedical NEs. To use it with the GeoNames, we manually tagged each place name to "LOC". The symbol "O" indicates that the word is not part of a geospatial NEs, and "LOC" indicates that the word is part of a geographical NEs.

III. RESULTS AND DISCUSSION

In addition to training and testing our system on Wiki and Geojo, we have conducted several tests on the training data set. We have used the following corpuses, Geojournal and Wikipedia, Reuters[8] (RC), CoNLL2003[9], GeoNames database to further train our corpus. As shown in Table 1, a combination of corpora has been used for this train. These combinations were tested under a YamCha tool.From the results shown in Table 1 the following remarks can be drawn.

Table 1: Results of Corpus combination

Corpus	Precision	Recall	F-measure
Geojo	0.50	0.63	0.56
Wiki	0.66	0.47	0.55
Geojo + Wiki	0.56	0.53	0.54
RC + CoNLL03 + GeoDB	0.50	0.89	0.64
GeoDB+RC+CoNLL03+GeoJo+wiki	0.60	0.80	0.69

Firstly, when using Geojo, we have found that the recall is higher than using Wiki'scorpus. On the other hand, the precision with Wiki is higher than the precision with Geojo. Secondly, when we have added Wiki to Geojo, we have got a balance in both the precision and the recall comparing to the individual results.Thirdly, when we used a corpus consisting of RC, CoNLL03, and GeoDB, we noticed that the recall is the highest result comparing with the other combination results, we have obtained 89%.The last but not the least, when we added (Geojo and Wiki) to the corpus of RC. CoNLL03, and GeoDB, we found that there is a trade-off between the recall and the precision comparing with the corpus consisting of RC, CoNLL03, and GeoDB. Thus, we have got the best performance F-measure 0.69%. Comparing with Witmer's work [10], the result of our F-measure is better than Witmer"s result. The latter has got 0.67%, while we have got 0.69%.

2. Evaluation

We are using CAFTETIER system[10] which is used to evaluate our results and perform looking-up in GeoNames database to separate actual place references and non-place references. From the comparison made between the results of YamCha and CAFETIER, as shown in Table 2, the following remarks can be drawn. Firstly, the precision is improved by %11 using CAFTETIER over YamCha. The main reason for this improvement is that CAFTETIER system contains a GeoNames database which helps the CAFTETIER system to extract the accurate entities. On the other hand, YamCha does not come with any geospatial database for the place names.Secondly, the recall result of CAFETIER system is less by %10 than the results of YamCha. This is because YamCha is given manually annotated data for any place names, regardlessits semantic meaning (i.e. Whether this place name refers to an organisation or a person's name, etc.). This allowed our system to retrieve a high volume of location names.Thirdly, the F-measure of both CAFETIRER and YamCha systems is approximately the same. This is due to the remark that the precision is increased and the recall is decreased in the CAFETIER results, whereas the precision was low and the recall was high in the YamCha results.

Table 2:Comarision between CAFETIER and YamCha

	Precision	Recall	F-Measure
YamCha Result	0.60	0.80	0.69
CAFETIER Result	0.71	0.70	0.67

CONCLUSIONS

The main aim of this project was to design and develop a software component based on the supervised machine learning algorithm to extract geographical named entities from the unstructured text. To achieve this aim, we have accomplished the following objectives. Firstly, we have investigated the approaches of supervised classification that could be used to deal with geographical entities in terms of both performance and accuracy. Thus, we have chosen the automatic training approach. This is because this approach does not need an expert

[7] Part-of-Speech tagging: the process of marking up the words in a text as corresponding to a particular part-of-speech, based on both its definition, as well as its context. See Appendix A.

[8] http://trec.nist.gov/data/reuters/reuters.html.

[9] http://www.cnts.ua.ac.be/conll2003/ner/

[10] http://www.nactem.ac.uk/cafetiere/

knowledgeengineer to annotate the text. Secondly, we havedesigned our system. This system consists of three stages. The first stage (i.e. Annotator); we have used a tool called GENIA tagger to annotate this corpus. Then, we manually annotated the GeoNames in the output of the GENIA tagger and chosen the word-features by which the GeoNames will be selected.In the second stage, called GeoExtractor, we have used a machine learning classifier, called YamCha, to train and to test our corpus.Then, YamCha is tested on a corpus of 60 articles taken from Wikipedia and Geojouranl. The overall results obtained are as follows: precision of 60% and a recall of 80% and F-Measure of 69%. Finally, we have evaluated the test results obtained from YamCha. To do so, we first examined a system called CAFETIERE and then used it to evaluate our results. This evaluation has shown approximately equivalent results to our obtained results.

ACKNOWLEDGMENTS

Author would like to express their deepest appreciation to Manchester university andculture bureau of saudiarabia in uk, which is supported by the government of the kingdom of Saudi Arabia.

REFERENCES

[1] Appelt, D. and Israel, D., 1999. Introduction to Information Extraction Technology: IJCAI-99 tutorial. [Online]Available at: <http://www.ai.sri.com/~appelt/ie-tutorial/IJCAI99.pdf> [Accessed 13 April 2011].

[2] J Tang, M Hong, D Zhang, B Liang, 2007. Information Extraction: Methodologies and Applications, pp.1-40.

[3] Mooney, R. J., Bunescu. R., 2005.Mining Knowledge from Text Using Information Extraction.ACM SIGKDD Explorations Newsletter. 7(1),pp.3-10.

[4] Zhou, G., Su, J., 2002. Named entity recognition using an HMM-based chunk tagger. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. July 2002 Philadelphia: USA pp.473-480.

[5] Lafferty, J., McCallum, A., Pereira, F.,2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. The 18th International Conference on Machine Learning.

[6] Li, W., McCallum, A., 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons.Proceedings of the 7th conference on natural language learning.

[7] McDonald, R., Pereira, F.,2005. Identifying gene and protein mentions in text using conditional random fields. BMC Bioinformatics. 6(1),pp.1-3.

[8] Christopher J. C. Burges, 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl.Discov., 2(2),pp.121–167.

[9] Pianta, Emanuele, Christian Girardi, and Roberto Zanoli. "The TextPro Tool Suite." LREC. 2008.

[10] Jeremy and JugalKalita, 2009. Extracting Geospatial Entities from Wikipedia. International Conference on Semantic Computing , pages 450 - 457 .

★★★