Statistical name detector

Matthias Gottlieb

Nordanvägen 3E: 003
22228 Lund
(Sweden)

Christoph-Probst-Str. 8 / 0918
80805 München
(Germany)

gottlieb@in.tum.de
matthias.gottlieb.583@student.lu.se

Abstract

This paper is about statistical name detection. This is a way you can generate information for different organizations for example newspapers. It helps to identify locations, organizations, persons and miscellaneous, which could not be directly attached to the three other categories. Over three different feature vectors the F-measure is increased up to 46.17%.

1 Introduction

This paper is analyzing a corpus and finding names which are common problems. This technique is useful for newspapers, collecting material for short information as well as messages and so on.

A name normally consists of parts which are very typical for a name. There are personal names like “Hans Eriksson”, “Max Müller Heisenberg” or “Christian von Steinhausen”. You can have one or more forenames, a title of nobility, a title (Professor, Dr., etc.) or aristocratically. There are also some typical indications for an organization like a share holding company that is ending with AG. I looked at an English corpus and found out that in the English language all names are capitalized.

A name is sorted in four different parts the LOC, MISC, ORG and PER, which stays for location, miscellaneous, organizations and persons.

First target was to get a state. This was given by the part of speech (POS) tag. This is analyzed by the J48-model given by weka. So the data consists of attributes, which are a name, a POS and a chunk tag. The chunk tag was replaced with the name tag and builds the baseline.

2 Rules

In order to find useful rules I decided to guess some rules and to find out how good they are working.

These rules are:

- look at POS,
- look for capitalized words,
- and look at suffixes of the words.

Suffixes are very useful even if you take a whole word they are still useful, because in the end of a name you will find some specific part for a name, for example, in the Swedish language a lot of names ending with “son”, or many countries ending with “land”. This helps to identify a name.

2.1 Baseline

The Baseline was built by the statistical chunk method which we used in our lab[1]. The chunk tags were substituted by the tagging of the names (PER, ORG, MISC, LOC, 0). Zero stays for no part of name. These works not directly because you have to make some small changes in the attributes, so weka is able to read it.

[1] Look at course webpage http://www.cs.lth.se/EDA171/cw3.shtml, called 10th of January 2009

2.2 Capital letters

In the English language all names are capitalized. The idea was to get all names by this method. But because of the POS where all names are marked as a noun I didn't expected an improvement. Solving these, a feature with true and false was built. This just went through all words and decided if it is capitalized or not and added the decision as a new row.

2.3 Suffixes of length two

At first I wanted to add every word but this leads to a problem with weka, because if the number of parts in an attribute is too much because weka needs a lot more memory than. In order to solve these build suffixes from each word and not longer than two. There are 24 letters per alphabet plus some extras for example U.S., where you have "S." as a suffix. So it is even a smaller number of arrangements instead of taking the whole word. But also by these numbers of parts you have to close some programs and give weka more memory to run it. In order to have a comparable result you also have to take all suffixes from the test corpus otherwise it could happen that it may not work.

3 Analysis and results

Surprising is that by adding a suffix the F-measure for locations and organization increased a lot.

I looked through 15.000 words from ca. 51.500 words long corpus. All words which had a different marking from the test set were marked and counted.

Just by running the program and having a lock on the POS the F-measure was 25.36% on the J48 model by using weka. This technique was similar to lab three from our course.[2]

Capitalization increased the percentage to 31.82% in the F-measure.

[2] Ibidem

The suffixes bring me up to 46.17% which was also more than expected. By adding these two small things getting near to 50% was more than ever thought.

For the three models rather the implemented features (Baseline, Capital letters and suffixes of length two) there were processed 51578 tokens with 5942 phrases.

3.1 Baseline

The Baseline found 5991 phrases, 1531 were correct, by an accuracy of 86.52%. The precision was 25.25% and in recall it found 25.46%.

By analysis more detailed you have no LOC and no MISC found. Some ORG by 3.97% what is absolutely unexpected and unexplainable, because there was just the POS tag so only PER should be marked, thus we got 38.71% for PER in the F-measure. The precision was 25.48% and in recall there where found 80.51%, The most PER were found by the recall. The ORG part was not expected; just because of completeness precision which was 17.54% and recall 2.24%.

3.2 Capital letters

The capital letters feature found 6610 phrases, 1997 were correct, by an accuracy of 87.62%. The precision was 30.21% and in recall it found 33.61%. The increasing of the accuracy by 1.1% shows how small this feature takes effect on the text. But an increasing of 6.46% in the F-measure is quite much.

What happened by doing this? The F-measure of MISC increases from zero to 51.99%. The precision was 63.68 and by recall I got 43.93%. This comes because there are some POS marked with NNP and some marked as NNS. The NNS is a MISC part of a name that is why it increases. These are for example American, Indian and so on. This leads to 31.82%, what is quite impressive because it was not expected, to have any effect with this.

Interesting the capitalization detects words for example American, but it produces failures. Countries and City names were not detected as location,

i.e. London is LOC or PER. Find first name mostly wrong, but the last name was often correctly found as PER.

City and team names are a huge problem to distinguish them, because for example the team name Birmingham is sometimes a city. But in the corpus this word was often a name of a team. There were often soccer results in the corpus where Birmingham was playing against another team, it is not the city meant who is playing. It is the soccer club Birmingham. This is heavy to detect statistical, when you don't have a look at the word before or after.

3.3 Suffixes of length two

The suffix feature found 7168 phrases, 3024 were correct, by an accuracy of 91.75%. The precision was 42.25% and in recall it found 50.89%. Instead of nearly 2000 words there are now more than 3000 words are correct identified. By recall there are more than 50% found which is amazing high for just two letters at the end of a word. Accuracy from more than 90% is sufficiently good and unexpected.

Furthermore the suffix feature helps obviously to learn to separate the name in the categories. That raise the F-measure for MISC to 57.57%, for LOC to 53.87%, for ORG to 34.51% and for PER to 43.51%.

First of all LOC has the highest growth this is because country names ends often with "nd" like Island, England, Poland, Ireland and some more. In the German language you have also the country name Germany which is spelled Deutschland in that language. Amassing is the precision by LOC which is 55.52%. More than the halves of the words are identified by the precision, which could have to do with the similar ending of the words.

The organization parts are in the last part (last word of the organization) mostly correctly identified this is because a lot of companies and organizations are ending with small words between one and four letters like GmbH, GBR, AG, AB, Ltd. and many more. The separation between a PER or LOC is much more difficult and needs further aspects to look at. As you can see the organization name in whole, for example Coca Cola GmbH & Co. KG. There you have the typically ending which is identified but the first part is wrong to keep it easier identified i.e. as PER. When we look behind CoCa we find Cola and after this the typical words which identifying the name as an organization. But nevertheless the precision increases to 31.63% and the recall increases to 37.96%.

By the MISC we got a small improvement by about 5% in the F-measure to 57.57%. The precision went a bit down from 63.68% to 61.35%, but the recall increases which produces the better F-measure. The recall went up from 43.93% to 54.23%. By looking through the corpus and comparing it with the test set I couldn't find any similarities in the endings which I had found by the country parts.

The persons are very interesting they increases from 40.92% to 43.51% that is not much in the F-measure, but the precision increases from 26.96% to 85.10%. That means there must be something which helps to identify them more precisely. In the Swedish language a lot of names are ending with "son" like Nilsson, Johansson, Eriksson, etc. In the corpus were a lot of football games where you find Swedish players but this could not have such a huge effect in total. So I had a look at the first names and recognized that a lot of first names are ending with "as" like Tobias, Matthias, Andreas, Thomas, etc. Only with the name of different writing of the name Matthias you get a lot of different person names the most common writings are Matthias or Mattias, but also you have Matias, Mathias. There are also a lot of names who are ending with "a" like Petra, Anna, Hanna, Andrea, Sofia, etc. It seems to be like a difference in the names between a female name and a male name. The recall went down from 84.80% to 57.22%. Suffixes helps to identify a person's name more precisely. By looking at the corpus I recognized that mostly the last name is identified correctly and not the first name. It seems that the suffix is not the best identification for the first name but a good indication. The examples which I showed up reflects in the ending, are markings which can help to identify a personal name.

To make a different between a city and a team name is not so easy and you need more than the

suffix. For a human being it is easy to separate it. You can hear it from the context. When we look some words before or after, in the sentence itself, we can find more precisely the organizations. But language is more difficult and some reflexive words are pointing to the sentences before. A lot of information can be found in small words like "the" and "of" which identify or connect the word to one part.

3.4 Comparison test corpus to train corpus

You can find in the tested train corpus that i.e. U.S. is correct tagged as LOC but U.S. Open is tagged as a MISC. So Open is tagged as a name which is right but not in the right category. The second addition looks at the last two letters so from U.S. it is "S." and from Open it is "en". What else is different of course, the POS part is tagged different.

Some wrong tagging of the test text were found for example Wimbledon was tagged as a MISC, but it was a LOC of the soccer game. Washington was correct tagged as a PER when it should be a person and also correctly tagged when it should be a location.

The longest found part of name was an organization. This was out of ten words.

4 Upgrades

For further improvements of the F-measure the PER and ORG can be mostly improved. To get a better PER result we have seen that in the analysis the forename is not always correct. To get rid of these have a look at one word before.

In order to have an improvement at a part of organizations, look at one or more words after the tested word. Organizations have special markings in different languages like GmbH, e.V., Ltd., AG, AB and so on. Also by results for example of football games you find behind the organization part a number or a double point. If you find one of these suggests that it could be an organization.

There are also some keywords on which we can have a look at for example you have a word like Open. Open can have different meanings. It could be a verb for example "please open the window" or a noun like "US Open". Than we have found a location and a MISC but in the context together it is one MISC. Other keywords are: The, League, Of, and so on. To get rid of this look at plus minus two words surround the tested word.

Furthermore a prefix could help, but you cannot have a good identification at the beginning because the words are starting too dissimilar.

5 Conclusion

The statistical name detection obviously works and can be improved by looking at some more details surround the tested word, could help. For testing in a sentence I expect that not more than plus / minus two words are necessary. Have also a look, if the word is having an article or not. And very important could be that the program learns prefixes and suffixes from a company by statistical detection. This leads us to a non word list where automatically it can find in different languages i.e. companies. These could be Ltd., GbR, GmbH, AG, AB and many more. Statistical name detection needs a huge corpus to improve itself, to be good enough to be used. Wordlist can help, but this helping is not very statistical. You can have a huge effect as this small experiment shows by adding small features.

Names can find statistically, analyzed and used for news papers and organizations who needs. To get absolute precision and extremely high accuracy a wordlist should be used, to get the performance which is needed. In 2003 the state of the art for precision was 88.99%, for recall 88.54% and the F-measure was 88.76+-0.7% reached by [FIJZ03].[3]

References

[FIJZ03]
Radu Florian, Abe Ittycheriah, Hongyan Jing and Tong Zhang, Named Entity Recognition through Classifier Combination. In: *Proceedings of CoNLL-2003*, Edmonton, Canada, 2003, pp. 168-171.

[3] Looked at http://www.cnts.ua.ac.be/conll2003/ner/, called 11th of January 2009

Further details
http://www.cnts.ua.ac.be/conll2003/ner/#FIJZ03, called 11th of January 2009

[Internet reference]
http://www.cnts.ua.ac.be/conll2003/ner/, called 11th of January 2009

[Internet reference]
http://www.cs.lth.se/EDA171/cw3.shtml, called 10th of January 2009