Text mining chemistry from Dutch or Swedish texts

Egon Willighagen ⬤

Citation

Keywords

Abstract

Oscar is a text miner. It mines in text for chemistry. Oscar4 is the next iteration of Oscar code that I worked on in the past three months, with Lezan, Sam, and David.

Copyright

chem-bla-ics

Oscar is a text miner. It mines in text for chemistry. Oscar4 is the next iteration of Oscar code that I worked on in the past three months, with Lezan, Sam, and David. I blogged about aspects of Oscar4 at several occasions:

- Working on Oscar for three months
- Oscar text mining in Taverna
- Multiple unit test inheritance with JExample
- Oscar4 Java API: chemical name dictionaries
- Oscar4 command line utilities
- Installing Oscar
- Adding a new dictionary to Oscar
- Status update on BJOC analysis with Oscar and ChemicalTagger
- Status update on BJOC analysis with Oscar and ChemicalTagger #2
- Supramolecular chemistry
- Status update on BJOC analysis with Oscar and ChemicalTagger #3
- Oscar: training data, models, etc

These posts will server is a some initial critical mass for a draft report I plan to finish today. I might have to blog some further posts with diagrams, here and there. This post is actually one of them, and discusses something where Oscar can be expected to go next, now that the design is cleaned up (though this effort is not halted now) and it has become possible again to extend it. The over 250 unit tests make this a lot easier too.

One aspect where I expect Oscar to go in 2011 is the support for other languages. To a very large extend this is based on multi-language support in the dictionaries, as well as having training data in a particular language. This also provides some context to my earlier post about the need for a Oscar training data repository .

This extension opens a number of options: analysis of patent literature in other languages, monitoring of press releases in other languages, and news items in local news papers, etc. For example, it could analyse this C2W news item on yeast cells:

chem-bla-ics

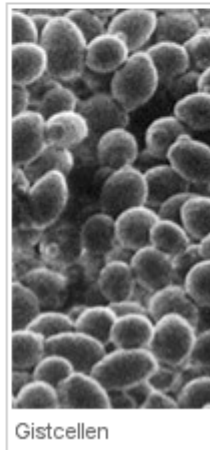There are many use cases for such localized text mining. And it surely matters for determining the impact of research.

Oscar has various places where language specifics are found. For example, in tokenization of a text. One step here is the detection of sentence ends. This is done in most western languages with a period, exclamation mark, question mark, etc. But periods (dots) are also used in abbreviations. Similarly, colons can be used in chemical names. But the every language comes in with different abbreviations that need to be recognized.

Currently, some abbreviations are found in NonSentenceEndings. In the past three months, we have been cleaning up the code, and restructured the source code, making it easier to detect such places. This class will likely undergo further refactoring, to making the list of such non-sentence-endings configurable via files or so. What I expect to see, is that we you initiate Oscar like this:

```
Oscar oscar = new Oscar(Locale.US);
```

This might actually even make a nice student summer project. The biggest challenge will be in making a good corpus of training data, like the SciBorg training data that was used for training Oscar3.

But the whole normalization is tainted with English language specifics too. For example, the normalizer will have to 'normalize' the question marks, for which there exist several unicode

chem-bla-ics

variations. But the normalized variant is language dependent. For example, greek and armenian have different characters (see this page), and then we have not even started talking about the right to left.

Besides localized dictionaries, this Oscar will also benefit from a localized OPSIN. It seem to recognize the Dutch propaan, but not benzeen. I am not going to look at that soon, but if you are interested, I recommend checking out Rich' posts about forking OPSIN and writing patches.

Getting Oscar going for other languages is a challenge, but also offers new opportunities. Just email the oscar mailing list if you are interested and need help.