chem-bla-ics

Oscar: training data, models, etc

Published December 26, 2010

Citation
Willighagen, E. (2010, December 26). Oscar: training data, models, etc. *Chem-bla-ics*. https://doi.org/10.59350/pa72q-ykk64

Keywords
Oscar, Textmining

Abstract

Oscar uses a Maximum Entropy Markov Model (MEMM) based on n-grams. Peter Corbett has written this up (doi:10.1186/1471-2105-9-S11-S4). So, it basically is statistics once more. If you really want a proper bioinformatics education, so do your PhD at a (proteo)chemometrics department.

Copyright

Oscar uses a Maximum Entropy Markov Model (MEMM) based on n-grams. Peter Corbett has written this up (doi:10.1186/1471-2105-9-S11-S4). So, it basically is statistics once more. If you really want a proper bioinformatics education, so do your PhD at a (proteo)chemometrics department.

N-grams are word parts of n characters. For example, the trigrams of acetic acid include `ace`, `cid`, `tic`, `eti`, and `aci`. N-grams of length four include acid, etic, and acet. The MEMM assigns weights to these n-grams, and based on that decided if something is in deed a *named entity* (in Oscar terminology). For example, consider the `acet` n-gram: acetone should be matched, but the n-gram `facet` not.

Put this in perspective in the ongoing refactoring of the Oscar software. We are changing normalization (e.g. converting all unicode hyphen alternatives into one specific hyphen), updating the tokenizer (e.g. changing the list of non-sentence-endings like *Prof.*). It is clear this changes the n-grams typical for chemical-like things. Worse, the weights are tuned towards to know n-grams, and statistical models are generally a bit overtrained for the data, or, at least, specific for it.

Now, if the distribution of n-grams changes, the weights in the model need to be updated too, to not degrade the model performance. So, Oscar is useless if we cannot retrain its MEMM component after a refactoring. If that would be impossible, we would have effectively created an *intellectual monopoly*.

Thus, what the Oscar project needs, is one or more free sets of annotated literature, which can be used to train new MEMM models. The SciBorg corpus was used to train the current Oscar3 and Oscar4 models. This data (copyright RSC) will very likely be available under a Creative Commons license (RSC++), but may have the NC clause, which would not be good for developing a business model around the opensource Oscar (such as providing a high-performance web service via a subscription service). I have recently written up the problems the NC clause introduces, and some examples of commercial Open Source cheminformatics projects.

We need not focus only on this SciBorg data, however. In fact, we will need multiple models anyway. For example, the SciBorg papers (42 if not mistaken) are around a particular kind of literature. So, it introduces the risk of using it to analyse papers out of the application domain. Furthermore, I am very interested (and others indicated so too) to use Oscar for other languages. Surely, English is the major language, but there are many use cases for Oscar when useful for other languages.

Therefore, for what we need in the Oscar project, is a registry of training (/test) data, annotated itself with metadata around how that data was created (what quality assurance, what kind of named entity types, how many domain experts were involved, etc), test results for those data sets, etc. My time on the Oscar project is almost over, and I have no clue when I will be able to invest the same amount of time into the project as I did in the past three months. But the creation of this registry is clear step that must be taken in the Oscar4 development.