NLP Seminar -Annotation of Scientific Abstracts

Nagesh Ramamoorthy -374559
RWTH AACHEN

Abstract

The following abstract acts as an example for end result of ”Annotation in scientific Abstracts”. The different colors represent different Annotations in abstracts

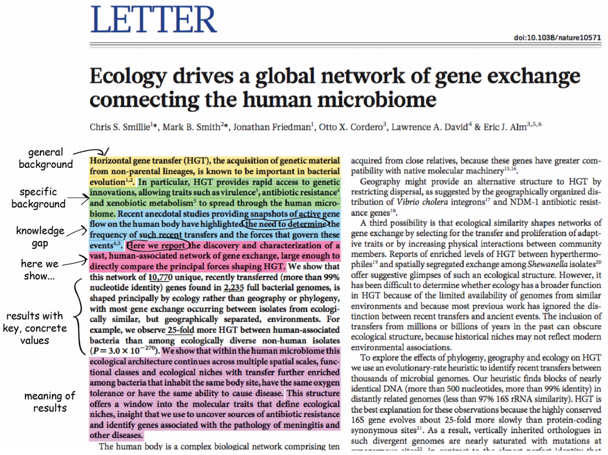
Aim: Requirement to access the specific type of information e.g. information about the objective,methods, results or conclusions has increased over years.**Background**: Several schemes have been developed to characterize such information in full journal papers. Yet many tasks focus on abstracts instead. **Method** : In this work we discuss and compare different techniques used in annotating a scientific article by considering the different features used for classification and labelling.Several methods including Naive Bayes, Support vector Machine , Conditional Random Field and Sequence Labelling used for annotation and classification are analyzed and compared through out the work .**Result**: CRF outperformed the SVM with features from previous and next sentence showing that is more adequate to classify sentences of scientific abstracts .**Conclusion**: Since features are mainly based on lexical contents of annotated text (unigrams and bigrams), the accuracy strongly improves when a greater dataset is considered.

1 Introduction

Recently the amount of scientific information available has increased at an unprecedented rate. Recent estimates reported that a new scientific paper is published every 20 seconds.Natural Language Processing Technology represents a key enabling factor in providing scientists with intelligent patterns to access to scientific information. Extracting information from scientific papers by annotation, for example, can contribute to the development of rich scientific knowledge bases which can be leveraged to support intelligent knowledge access and question answering.

Scientific abstracts tend to be very similar in terms of their information structure. For example, many abstracts provide some background information before defining the precise objective of the study,and the conclusions are typically preceded by the description of the results obtained.Many readers of scientific abstracts are interested in specific types of information only, e.g.the general background of the study, the methods use in the study, or the results obtained . In recent years, numerous initiatives have emerged to automatically process electronic documents in the life sciences, add semantic markup to them and facilitate access to scientific facts. To date, a number of different schemes and techniques have been proposed for sentence-based classification of scientific literature according to information structure, e.g. (Teufel and Moens, 2002; Mizuta et al., 2005; Lin et al., 2006; Hirohata et al., 2008; Teufel et al., 2009; Shatkay et al., 2008; Liakata et al., 2010).

In this work we investigate annotation based on three schemes - – those based on Section Names (S1), Argumentative Zones (S2) and Core Scientific Concepts (S3): S1 : It is based on section names found in some scientific abstracts. We use the 4-way classification from (Hirohata et al., 2008) where abstracts are divided into objective, method, results and conclusions. S2 : Argumentative Zoning Annotation Schema bundles together similar rhetorical moves casting the general argumentation recognition Knowledge Claim Discourse Model into a sentence classification task S3 : CoreSC has more granularity when dealing with content-relate categories. Fig.1 shows the idea behind the annotation in scientific articles.



LETTER

doi:10.1038/nature10571

Ecology drives a global network of gene exchange connecting the human microbiome

Chris S. Smillie[1]*, Mark B. Smith[2]*, Jonathan Friedman[1], Otto X. Cordero[3], Lawrence A. David[4] & Eric J. Alm[3,5,6]

Horizontal gene transfer (HGT), the acquisition of genetic material from non-parental lineages, is known to be important in bacterial evolution[1,2]. In particular, HGT provides rapid access to genetic innovations, allowing traits such as virulence[3], antibiotic resistance[4] and xenobiotic metabolism[5] to spread through the human microbiome. Recent anecdotal studies providing snapshots of active gene flow on the human body have highlighted the need to determine the frequency of such recent transfers and the forces that govern these events[4,5]. Here we report the discovery and characterization of a vast, human-associated network of gene exchange, large enough to directly compare the principal forces shaping HGT. We show that this network of 10,770 unique, recently transferred (more than 99% nucleotide identity) genes found in 2,235 full bacterial genomes, is shaped principally by ecology rather than geography or phylogeny, with most gene exchange occurring between isolates from ecologically similar, but geographically separated, environments. For example, we observe 25-fold more HGT between human-associated bacteria than among ecologically diverse non-human isolates ($P = 3.0 \times 10^{-270}$). We show that within the human microbiome this ecological architecture continues across multiple spatial scales, functional classes and ecological niches with transfer further enriched among bacteria that inhabit the same body site, have the same oxygen tolerance or have the same ability to cause disease. This structure offers a window into the molecular traits that define ecological niches, insight that we use to uncover sources of antibiotic resistance and identify genes associated with the pathology of meningitis and other diseases.

The human body is a complex biological network comprising ten

acquired from close relatives, because these genes have greater compatibility with native molecular machinery[15,16].

Geography might provide an alternative structure to HGT by restricting dispersal, as suggested by the geographically organized distribution of *Vibrio cholera* integrons[17] and NDM-1 antibiotic resistance genes[18].

A third possibility is that ecological similarity shapes networks of gene exchange by selecting for the transfer and proliferation of adaptive traits or by increasing physical interactions between community members. Reports of enriched levels of HGT between hyperthermophiles[19] and spatially segregated exchange among *Shewanella* isolates[20] offer suggestive glimpses of such an ecological structure. However, it has been difficult to determine whether ecology has a broader function in HGT because of the limited availability of genomes from similar environments and because most previous work has ignored the distinction between recent transfers and ancient events. The inclusion of transfers from millions or billions of years in the past can obscure ecological structure, because historical niches may not reflect modern environmental associations.

To explore the effects of phylogeny, geography and ecology on HGT we use an evolutionary-rate heuristic to identify recent transfers between thousands of microbial genomes. Our heuristic finds blocks of nearly identical DNA (more than 500 nucleotides, more than 99% identity) in distantly related genomes (less than 97% 16S rRNA similarity). HGT is the best explanation for these observations because the highly conserved 16S gene evolves about 25-fold more slowly than protein-coding synonymous sites[21]. As a result, vertically inherited orthologues in such divergent genomes are nearly saturated with mutations at synonymous sites[22], in contrast to the almost perfect identity that

Figure 1: Annotated Scientific Article

2 Data

For single label classification the classical ART dataset was considered. ART Corpus consists of 265 papers from the domains of chemistry an biochemistry . The 265 paper aggregates to 39,915 sentences.The corpus is available online at https://www.aber.ac.uk/en/cs/research/cb/projects/art/artcorpus/.

Multi-Core SC corpus dataset is a prominent dataset for Multi-Label classification.The corpus consists of 50 journal papers from the discipline of cancer risk assessment. Each of the 50 papers is annotated at the sentence level with at least one CoreSC (to a maximum of three). The 50 papers correspond to 8,501 unique sentences with sentence count per paper ranging between 85 and 432.The corpus is available online at http://www.sapientaproject.com/wp-content/uploads/2016/05/consensusannotate .zip.

3 Features

The first step in automatic identification of information structure is feature extraction. The work chooses a number of general purpose features suitable for all the three schemes. With the exception of the novel verb class feature, the features are similar to those employed in related works, e.g. (Teufel and Moens, 2002; Mullen et al., 2005; Hirohata et al., 2008): The following are all implemented as binary features:

- History: There are typical patterns in the information structure, e.g. RES tends to be followed by CON rather than by BKG.

- Absolute Location:Work divides the document into 10 unequal segments (as in Loc of (Teufel, 2000)) and assign 1 of the 10 locations, A–J, to the sentences. Larger segments, containing more sentences, are designated to be in the middle of the paper.

- Word:Like many text classification tasks, consider all the words in the present corpus as features.

- Bi-gram: Considered each bi-gram (combination of two word features) as a feature.

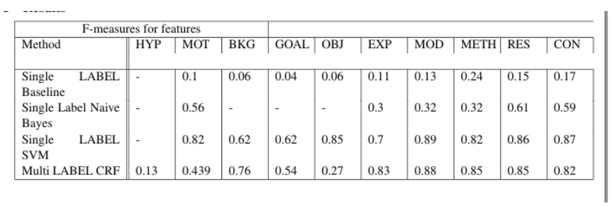
- Verb POS (VPOS): For each verb within the sentence , determine which of the six binary POS tags (VBD, VBN, VBG, VBZ, VBP and VB) representing the tense, aspect and person of a verb are present.

- POS:Tense tends to vary from one category to another, e.g. past is common in RES and past partici103 ple in CON. Paper used the part-of-speech (POS) tag of each verb assigned by the CC tagger (Curran et al., 2007) as a feature.

- GR:Structural information about heads and dependents has proved useful in text classification. The paper used grammatical relations (GRs) returned by the CC parser as features. They consist of a named relation, a head and a dependent, and possibly extra parameters depending on the relation involved, e.g. (dobj investigate mouse). Authors created features for each subject (ncsubj), direct object (dobj), indirect object (iobj) and second object (obj2) relation in the corpus.

4 Methods

- Single Label-NB and SVM : The work used Naive Bayes (NB) and Support Vector Machines (SVM) for classification. NB is a simple and fast method while SVM has yielded high performance in many text classification tasks.
NB applies Bayes' rule and Maximum Likelihood estimation with strong independence assumptions. It aims to select the class c with maximum probability given the feature set F . SVM constructs hyperplanes in a multidimensional space that separates data points of different classes. Good separation is achieved



F-measures for features										
Method	HYP	MOT	BKG	GOAL	OBJ	EXP	MOD	METH	RES	CON
Single LABEL Baseline	-	0.1	0.06	0.04	0.06	0.11	0.13	0.24	0.15	0.17
Single Label Naive Bayes	-	0.56	-	-	-	0.3	0.32	0.32	0.61	0.59
Single LABEL SVM	-	0.82	0.62	0.62	0.85	0.7	0.89	0.82	0.86	0.87
Multi LABEL CRF	0.13	0.439	0.76	0.54	0.27	0.83	0.88	0.85	0.85	0.82

Figure 2: Summarization of F-Measure

by the hyperplane that has the largest distance from the nearest data points of any class.

- Improved Features and Multi Label -CRF and SVM : The work used the trained Conditional Random Field (CRF) model .Conditional Random Fields are a Discriminative model, and their underlying principle is that they apply Logistic Regression on sequential inputs.

5 Results -Summary

Figure 2 gives the summarization of results from different methods used for annotation. SVM uncovers as many as nine of the 11 categories with accuracy of 81%. Six categories perform well, with F-measure higher than 60. EXP, BKG and GOAL have F-measure of 70, 62 and 62, respectively. Like the missing category HYP , GOAL is very low in frequency. The lower performance of the higher frequency EXP and BKG is probably due to low precision in distinguishing between EXP and METH, and BKG and other categories, respectively.

For Multi-Label - 12.5 percent of sentences obtained a multiCoreSC label .Most influential features of CoreSC annotation are domain specific.The best category recognised is still ”Method”, which is probably unsurprising as this is the largest group of annotations in the CRA corpus . While the categories ’Con’, ’Goa’, ’Mot’, and ’Obs’ are all comparatively small subsets of the overall corpus (between 2.1% and 7.4%), the model achieves respectable performance measures for this group: F-measures all in the range of 43% to 54.2%. Especially, the Goal category seems to be sufficiently concise as it constitutes only 2.1% of the gold standard but still yields an F-measure of 54.2%.

6 Conclusion

If we merge CoreSC categories so that we consider a coarser grain layer of four categories, namely Prior (BAC), Approach (MET+MOD+EXP), Outcome (OBS+RES+CON) and Objective (MOT+GOA+HYP+OBJT) then our F-measures respectively become: BAC: 59%, Approach: 72%, Outcome: 81%, Objective: 38%. A variant merge with seven categories, roughly corresponding to the scheme proposed by de Waard et al., 2009, which considers BAC, HYP, Problem(=MOT), GOA=(GOA+OBJT), MET= (MET+EXP+MOD), RES=(OBS+RES), Implication(=CON), gives us F1: BAC: 60%, CON: 44%, MET: 72%, GOA: 47%, MOT: 19%, HYP: 18% and RES: 72% .

There is not always a direct correlation of annotator agreement and classifier performance: Experiment and Model have an higher F-score but low interannotator agreement.

7 Improvements

Similar features and methods tend to perform the best / worst for each of the schemes. It is therefore unlikely that considerable scheme specific tuning will be necessary. However, developing new features further is necessary to make better use of the sequential nature of information structure. The different types of conceptualization zones defined by CoreSCs (Background, Hypothesis, Method, etc.) so far have been used to create extractive summaries and more use cases of filtering text during information extraction are in progress. Work in progress also involves the application of CoreSC annotations to full papers

References

[Maria Liakata et.al.2011] ufan Guo,Anna Korhonen,Maria Liakata and Ilona Silins. 2011. *Identifying the Information Structure of Scientific Abstracts: An Investigation of Three Different Schemes. Cambridge University Press, Cambridge, UK.*

[Maria Liakata et.al.2016] aria Lakata,Anika Oellrich, Shyamasree Saha. 2015. Multi-label Annotation in Scientific Articles - The Multi-label Cancer Risk Assessment Corpus. Cambridge University Press, Cambridge, UK.

[Maria Liakata et.al.2012] aria Lakata,Anika Oellrich, Shyamasree Saha. 2012. *Automatic recognition of conceptualization zones in scientific articles and two life science applications Cambridge University Press, Cambridge, UK.*