Social and Semantic Diversity: Socio-semantic Representation of a Scientific Corpus

Elisa Omodei
LATTICE and ISC-PIF
CNRS & ENS & U. Sorbonne Nouvelle
1 rue Mauriece Arnoux
92120 Montrouge France
elisa.omodei@ens.fr

Yufan Guo
University of Washington
Computer Science
Engineering
Box 352350 Seattle, WA 98195-2350
yufanguo@cs.washington.edu

Jean-Philippe Cointet
INRA Sens and ISC-PIF
Cité Descartes, 5 boulevard Descartes
77454 Marne-la-Vallée Cedex France
75013 Paris France
jphcoi@yahoo.fr

Thierry Poibeau
LATTICE
CNRS & ENS & U. Sorbonne Nouvelle
1 rue Mauriece Arnoux
92120 Montrouge France
thierry.poibeau@ens.fr

Abstract

We propose a new method to extract keywords from texts and categorize these keywords according to their informational value, derived from the analysis of the argumentative goal of the sentences they appear in. The method is applied to the ACL Anthology corpus, containing papers on the computational linguistic domain published between 1980 and 2008. We show that our approach allows to highlight interesting facts concerning the evolution of the topics and methods used in computational linguistics.

1 Introduction

Big data makes it possible to observe in vivo the dynamics of a large number of different domains. It is particularly the case in the scientific field, where researchers produce a prolific literature but also other kinds of data like numbers, figures, images and so on. For a number of domains, large scientific archives are now available over several decades.

This is for example the case for computational linguistics. The ACL Anthology contains more than 24,500 papers, for the most part in PDF format. The oldest ones date back to 1965 (first edition of the COLING conference) but it is mostly after 1980 that data are available in large volumes so that they can be exploited in evolution studies. The volume of data increases over time, which means there is a wide diversity in the number of papers available depending on the given period of time. There are similar archives for different domains like, e.g. physics (the APS database provided by the American Physical Society) or the bio-medical domain (with Medline).

These scientific archives have already given birth to a large number of different pieces of work. Collaboration networks have for example been automatically extracted so as to study the topology of the domain (Girvan and Newman, 2002) or its morphogenesis (Guimera et al., 2005). Referencing has also been the subject of numerous studies on inter-citation (Garfield, 1972) and co-citation (Small, 1973). Other variables can be taken into account like the nationality of the authors, the projects they are involved in or the research institutions they belong to, but it is the analysis of the textual content (mostly titles, abstracts and keywords provided with the papers) that have attracted the most part of the research in the area since the seminal work of Callon (Callon et al., 1986; Callon et al., 1991).

In this paper, our goal is to investigate the evolution of the field of computational linguistics, which means that text will play a crucial role. Textual analysis is then mixed with the study of individual trajectories in the semantic space: our goal is to propose possible avenues for the study of the dynamics of innovation in the computational lin-

Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014, pages 71–79,
Gothenburg, Sweden, April 26 2014.

guistics domain.

The ACL Anthology has been the subject of several studies in 2012, for the 50 years of the ACL. More specifically, a workshop called "Rediscovering 50 Years of Discoveries" was organized to examine 50 years of research in NLP (but, for the reasons given above, the workshop mostly focused on the evolution of the domain since 1980). This workshop was also an opportunity to study a large scientific collection with recent NLP techniques and see how these techniques can be applied to study the dynamics of a scientific domain.

The analysis of this kind of data is generally based on the extraction of key information (authors, keywords) and the discovery of their relationships. The data can be represented as a graph, therefore graph algorithmics can be used to study the topology and the evolution of the graph of collaborations or the graph of linked authors. It is thus possible to observe the evolution of the domain, check some hypotheses or common assumptions about this evolution and provide a strong empirical basis to epistemology studies.

The paper "Towards a computational History of the ACL: 1980-2008" is very relevant from this point of view (Anderson et al., 2012). The authors try to determine the evolution of the main sub-domains of research within NLP since 1980 and they obtain very interesting results. For example, they show the influence of the American evaluation campaigns on the domain: when a US agency sponsored a sub-domain of NLP, one can observe a quick concentration effect since a wide number of research groups suddenly concentrated their efforts on the topic; when no evaluation campaign was organized, research was much more widespread across the different sub-domains of NLP. Even if this is partially predictable, it was not obvious to be able to show this in a collection of papers as large as the ACL Anthology.

Our study has been profoundly influenced by the study by Anderson et al. However, our goal here is to characterize automatically the keywords based on the information they carry. We will thus combine keyword extraction with text zoning so as to categorize the keywords depending on their context of use.

The rest of the paper is organized as follows. We first present an analysis of the structure of abstracts so as to better characterize their content by mixing keyword extraction with text zoning. We show how these techniques can be applied to the ACL Anthology in order to examine specific facts, more specifically concerning the evolution of the techniques used in the computational linguistics domain.

2 A Text Zoning Analysis of the ACL Anthology

The study of the evolution of topics in large corpora is usually done through keyword extraction. This is also our goal, but we would like to be able to better characterize these keywords and make a difference, for example, between keywords referring to concepts and keywords referring to methods. Hence, the context of these keywords seems highly important. Consequently, we propose to use Text Zoning that can provide an accurate characterization of the argumentative goal of each sentence in a scientific abstract.

2.1 Previous work

The first important contributions in text zoning are probably the experiments by S. Teufel who proposed to categorize sentences in scientific papers (and more specifically, in the NLP domain) according to different categories (Teufel, 1999) like BKG: General scientific background, AIM: Statements of the particular aim of the current paper or CTR: Contrastive or comparative statements about other work. This task is called Rhetorical zoning or Argumentative zoning since the goal is to identify the rhetoric or argumentative role of each sentence of the text.

The initial work of Teufel was based on the manual annotation of 80 papers representing the different areas of NLP (the corpus was made of papers published within the ACL conferences or Computational Linguistics). A classifier was then trained on this manually annotated corpus. The author reported interesting results despite "a 20% diference between [the] system and human performance" (Teufel and Moens, 2002). The learning method used a Naive Bayesian model since more sophisticated methods tested by the author did not obtain better results. Teufel in subsequent publications showed that the technique can be used to produce high quality summaries (Teufel and Moens, 2002) or precisely characterize the different citations in a paper (Ritchie et al., 2008).

The seminal work of Teufel has since then given rise to different kinds of works, on the one hand to refine the annotation method, and on the other hand to check its applicability to different scientific domains. Concerning the first point, research has focused on the identification of relevant features for classification, on the evaluation of different learning algorithms for the task and more importantly on the reduction of the volume of text to be annotated. Concerning the second point, it is mostly the biological and bio-medical domains that have attracted attention, since scientists in these domains often have to access the literature "vertically" (i.e. experts may need to have access to all the methods and protocols that have been used in a specific domain) (Mizuta et al., 2006; Tbahriti et al., 2006).

Guo has since developed a similar trend of research to extend the initial work of Teufel (Guo et al., 2011; Guo et al., 2013): she has tested a large list of features to analyze the zones, evaluated different learning algorithms for the task and proposed new methods to decrease the number of texts to be annotated. The features used for learning are of three categories: *i*) positional (location of the sentence inside the paper), *ii*) lexical (words, classes of words, bigrams, etc. are taken into consideration) and *iii*) syntactic (the different syntactic relations as well as the class of words appearing in subject or object positions are taken into account). The analysis is thus based on more features than in Teufel's initial work and requires a parser.

2.2 Application to the ACL Anthology corpus

In our experiment, we only used the abstracts of the papers. Our hypothesis is that abstracts contain enough information and are redundant enough to study the evolution of the domain. Taking into consideration the full text would probably give too many details and thus introduce noise in the analysis.

The annotation scheme includes five different categories, which are the following: OBJECTIVE (objectives of the paper), METHOD (methods used in the paper), RESULTS (main results), CONCLUSION (conclusion of the paper), BACKGROUND (general context), as in (Reichart and Korhonen, 2012). These categories are also close to those of (Mizuta et al., 2006; Guo et al., 2011; Guo et al., 2013) and have been adapted to abstracts (as opposed to full text[1]). It seems relevant to take into consideration an annotation scheme that has already been used by various authors so that the results are easy to compare to others.

Around one hundred abstracts from the ACL Anthology have then been manually annotated using this scheme ($\sim$500 sentences; ACL abstracts are generally quite short since most of them are related to conference papers). The selection of the abstracts has been done using stratified sampling over time and journals, so as to obtain a representative corpus (papers must be related to different periods of time and different sub-areas of the domain). The annotation has been done according to the annotation guideline defined by Y. Guo, especially for long sentences when more than one category could be applied (preferences are defined to solve complex cases[2]).

The algorithm defined by (Guo et al., 2011) is then adapted to our corpus. The analysis is based on positional, lexical and syntactic features, as explained above. No domain specific information was added, which makes the whole process easy to reproduce. As for parsing, we used the C&C parser (James Curran and Stephen Clark and Johan Bos, 2007). All the implementation details can be found in (Guo et al., 2011), especially concerning annotation and the learning algorithm. As a result, each sentence is associated with a tag corresponding to one of the zones defined in the annotation scheme.

2.3 Results and Discussion

In order to evaluate the text zoning task, a number of abstracts were chosen randomly ($\sim$300 sentences that do not overlap with the training set). CONCLUSION represented less than 3% of the sentences and was then dropped for the rest of the analysis. The four remaining zones are unequaly represented: 18.05 % of the sentences refer to BACKGROUND, 14.35% to OBJECTIVE, 14.81 % to RESULT and 52.77 % to METHOD. Just by looking at these numbers, one can see how

[1]The categories used in (Teufel, 1999) were not relevant since this model focused on full text papers, with a special emphasis on the novelty of the author's work and the attitude towards other people's work, which is not the case here.

[2]The task is to assign the sentence only a single category. The choice of the category should be made according to the following priority list: Conclusion > Objective > Result > Method > Background. The only exception is that when 75% or more of the sentence belongs to a less preferred category, then that category will be assigned to the sentence.

Table 1: Result of the text zoning analysis (precision)

Category	Precision
Objective	83,87 %
Background	81,25 %
Method	71,05 %
Results	82,05 %

Figure 1: An abstract annotated with text zoning information. Categories are indicated in bold face.

Most of errors in Korean morphological analysis and POS (Part-of-Speech) tagging are caused by unknown morphemes . **BACKGROUND**
This paper presents a generalized unknown morpheme handling method with POSTAG(POStech TAGger) which is a statistical/rule based hybrid POS tagging system . **OBJECTIVE**
The generalized unknown morpheme guessing is based on a combination of a morpheme pattern dictionary which encodes general lexical patterns of Korean morphemes with a posteriori syllable tri-gram estimation . **METHOD**
The syllable tri-grams help to calculate lexical probabilities of the unknown morphemes and are utilized to search the best tagging result . **METHOD**
In our scheme , we can guess the POS's of unknown morphemes regardless of their numbers and positions in an eojeol , which was not possible before in Korean tagging systems . **RESULTS**
In a series of experiments using three different domain corpora , we can achieve 97% tagging accuracy regardless of many unknown morphemes in test corpora . **RESULTS**

methodological issues are important for the domain.

We then calculate for each of the categories, the percentage of sentences that received the right label, which allows us to calculate precision. The results are given in table 1.

These results are similar to the state of the art (Guo et al., 2011), which is positive taking into consideration the small number of sentences annotated for training. The diversity of the features used makes it easy to transfer the technique from one domain to the other without any heavy annotation phase. Results are slightly worse for the METHOD category, probably because this category is more diverse and thus more difficult to recognize. The fact that NLP terms can refer either to objectives or to methods also contributes rendering the recognition of this category more difficult.

Figure 1 shows an abstract annotated by the text zoning module (the paper is (Lee et al., 2002): it has been chosen randomly between those containing the different types of zones). One category is associated with each sentence but this is sometimes problematic: for example the fact that a hybrid method is used is mentioned in a sentence that is globally tagged as OBJECTIVE by the system. However, sentences tagged as METHOD contain relevant keywords like *lexical pattern* or *tri-gram estimation*, which makes it possible to infer that the approach is hybrid. One can also spot some problems with digitization, which are typical of this corpus: the ACL Anthology contains automatically converted files to PDF, which means texts are not perfect and may contain some digitization errors.

3 Contribution to the Study of the Evolution ACL Anthology

As said above, we are largely inspired by (Anderson et al., 2012). We think the ACL Anthology is typical since it contains papers spanning over more than 30 years: it is thus interesting to use it as a way to study the main evolutions of the computational linguistics domain. The method can of course also be applied to other scientific corpora.

3.1 Keyword extraction and characterization

The first step consists in identifying the main keywords of the domain. We then want to more precisely categorize these keywords so as to identify the ones specifically referring to methods for example. From this perspective, keywords appearing in the METHOD sections are thus particularly interesting for us. However, one major problem is that there is no clear-cut difference between goals and methods in NLP since most systems are made of different layers and require various NLP techniques. For example, a semantic analyzer may use a part-of-speech tagger and a parser, which means NLP tools can appear as part of the method.

Keyword extraction aims at automatically extracting relevant keywords from a collection of texts. A popular approach consists in first extracting typical sequences of tags that are then filtered according to specific criteria (these criteria can include the use of external resources but they are more generally based on scores mixing frequency and specificity (Bourigault and Jacquemin, 1999; Frantzi and Ananiadou, 2000)). In this study, we voluntarily used a minimal approach for keyword extraction and filtering since we want to keep most

Table 2: Most specific keywords found in the METHOD sections.

Methods		
Category	Method	N-grams
Machine learning	Bayesian methods	baesyan
	Vector Space model	space model, vector space, cosine
	Genetic algorithms	genetic algorithms
	HMM	hidden markov models, markov model
	CRF	conditional random fields
	SVM	support vector machines
	MaxEnt	maximum entropy model, maximum entropy approach, maximum entropy
	Clustering	clustering algorithm, clustering method, word clusters, classification problem
Speech & Mach. Trans.	Language models	large-vocabulary, n-gram language model, Viterbi
	Parallel Corpora	parallel corpus, bilingual corpus, phrase pairs, source and target languages, sentence pairs, word pairs, source sentence
	Alignment	phrase alignment, alignment algorithm, alignment models, ibm model, phrase translation, translation candidates, sentence alignment
NLP Methods	POS tagging	part-of-speech tagger, part-of-speech tags
	Morphology	two-level morphology, morphological analyzer, morphological rules
	FST	finite-state transducers, regular expressions, state automata, rule-based approach
	Syntax	syntactic categories, syntactic patterns, extraction patterns
	Dependency parsing	dependency parser, dependency graphs, prague dependency, dependency treebank, derivation trees, parse trees
	Parsing	grammar rules, parser output, parsing process, parsed sentences, transfer rules
	Semantics	logical forms, inference rules, generative lexicon, lexical rules, lexico-syntactic, predicate argument
Applications	IE and IR	entity recognition, answer candidates, temporal information, web search, query expansion, google, user queries, keywords, query terms, term recognition
	Discourse	generation component, dialogue acts, centering theory, lexical chains, resolution algorithm, generation process, discourse model, lexical choice
	Segmentation	machine transliteration, phonological rules, segmentation algorithm, word boundaries
Words and Resource	Lexical knowledge bases	lexical knowledge base, semantic network, machine readable dictionaries, eurowordnet, lexical entries, dictionary entries, lexical units, representation structures, lookup
	Word similarity	word associations, mutual information, semantic relationships, word similarity, semantic similarity, semeval-2007, word co-occurrence, synonymy
	Corpora	brown corpus, dialogue corpus, annotation scheme, tagged corpus
Evaluation	Evaluation	score, gold standard, evaluation measures, estimation method
Calculation & complexity	Software	tool development, polynomial time, software tools, series of experiments, system architecture, runtime, programming language
	Constraints	relaxation, constraint satisfaction, semantic constraints

of the information for the subsequent text zoning phase. We thus used NLTK for part-of-speech tagging and from this result extracted the most common noun phrases. We used a pre-defined set of grammatical patterns to extract noun phrases defined as sequences of simple sequences (e.g. adjectives + nouns, “phrase pairs”, “dependency graph”, etc.) possibly connected to other such patterns through propositions to form longer phrases (e.g. ”series of experiments”). Only the noun phrases appearing in more than 10 papers are kept for subsequent processing.

Candidate keywords are then ranked per zone, according to their specificity (the zone they are the most specific of) . Specificity corresponds to the Kolmogorov-Smirnov test that quantifies a distance between the empirical distribution functions of two samples. The test is calculated as follows:

$$D = \max_x |S_{N_1}(x) - S_{N_2}(x)| \quad (1)$$

where $S_{N_1}(x)$ et $S_{N_2}(x)$ are the empirical distribution function of the two samples (that correspond in our case to the number of occurrences of the keyword in a given zone, and to the total number of occurrences of all the keywords in the same zone, respectively) (Press et al., 2007). A high value of D for a given keyword means that it is highly specific of the considered zone. At the opposite, a low value means that the keyword is spread over the different zones and not really specific of any zone.

The first keywords of each category are then categorized by an expert of the domain. For the METHOD category, we obtain Table 2. Logically, given our approach, the table does not contain all the keywords relevant for the computational linguistics domain, but it contains the mots specific ones according to the above approach. One should thus not be surprised not to see all the keywords used in the domain.

3.2 Evolution of methods over time

The automatic analysis of the corpus allows us to track the main evolutions of the field over time. During the last 30 years, the methods used have changed to a large extent, the most notable fact being probably the generalization of machine learning methods since the late 1990s. This is outlined by the fact that papers in the domain nowadays nearly always include a section that describes an experiment and some results.

To confirm this hypothesis, we observe the relative frequency of sentences tagged as RESULTS in the papers over time. In the figure 3, we see that the curve increases almost linearly from the early 1980s until the late 2000s.

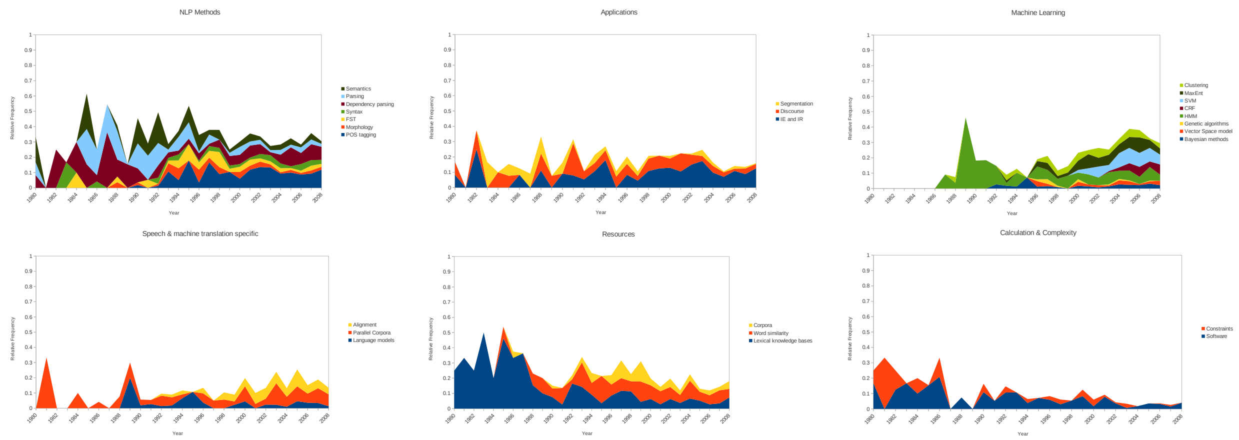

Figure 2: Evolution of the relative frequency of the different groups of methods over time.

It is also possible to make more fine-grained observations, for example to follow over time the different kinds of methods under consideration. The results are shown in figure 2. Rule based methods and manually crafted resources are used all over the period, while machine learning based methods are more and more successful after the late 1990s. This is not surprising since we know that machine learning is now highly popular within the field. However, symbolic methods are still used, sometimes in conjunction with learning methods. The two kinds of methods are thus more complementary than antagonistic.

One could observe details that should be checked through a more thorough study. We observe for example the success of dependency parsing in the end of the 1980s (probably due to the success of the Tree Adjoining Grammars at the time) and the new popularity of this area of research in the early 2000s (dependency parsing has been the subject of several evaluation campaigns in the 2000s, see for example for the CONLL shared tasks from 2006 to 2009).

Different machine learning methods have been popular over time but each of them continues to be used after a first wave corresponding to their initial success. Hidden Markov Models and n-grams are highly popular in the 1990s, probably thanks to the experiments made by Jelinek and his colleagues, which will open the field of statistical machine translation (Brown et al., 1990). SVM and CRF have had a more recent success as everybody knows.

We are also interested in the distribution of these methods between papers and authors. Figure 4 shows the average number of keywords

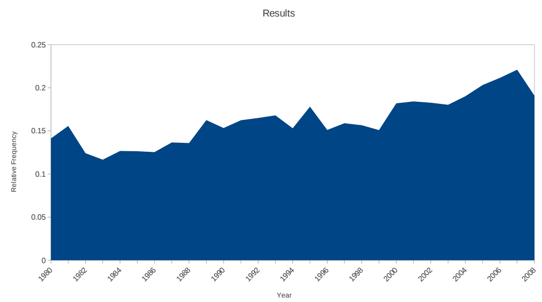

Figure 3: Evolution of the relative frequency of sentences tagged as RESULTS in the abstracts of the papers

appearing in the METHOD section of the papers over time. We see that this number regularly increases, especially during the 1980s, showing possibly a gradually increasing complexity of the systems under consideration.

Lastly, figure 5 shows the number of authors who are specialists of one or several methods. Most of the authors just mention one method in their papers and, logically, the curves decrease, which means that there are few authors who are really specialists of many methods. This result should be confirmed by a larger scale study taking into account a larger number of keywords but the trend seems however interesting.

3.3 The dynamics of the authors in the method space

One could say that the results we have reported in the previous section are not new but rather confirm some already well known facts. Our method allows to go one step further and try to answer more

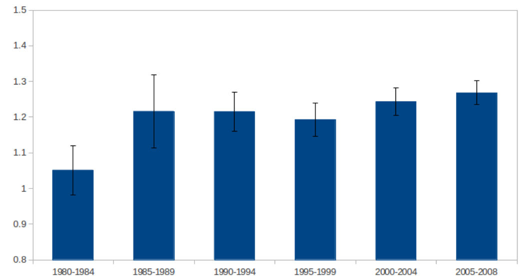

Figure 4: Evolution of the number of keywords related to methods over time.

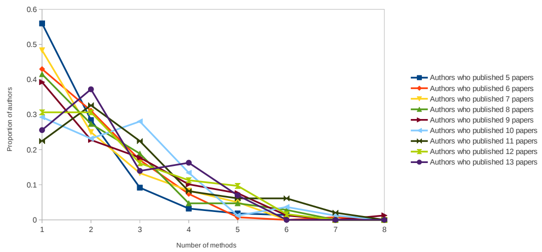

Figure 5: Proportion of authors specialized in a given number of methods (i.e. mentioning frequently the name of the method in the abstracts), for different categories of researchers.

challenging questions. How are new methods introduced in the field? Are they mainly brought by young researchers or is it mainly confirmed researchers who develop new techniques (or import them from related fields)? Are NLP experts specialized in one field or in a wide variety of different fields?

These questions are of course quite complex. Each individual has his own expertise and his own history but we think that automatic methods can provide some interesting trends over time. For example, (Anderson et al., 2012) show that evaluation campaigns have played a central role at certain periods of time, which does not mean of course that there was no independent research outside these campaigns at the time. Our goal is thus to exhibit some tendencies that could be interpreted or even make it possible to compare the evolution of the computational linguistics field with other fields. Out tools provide some hypotheses that must of course be confirmed by further observations and analysis. We do not claim that they provide an exact and accurate view of the domain.

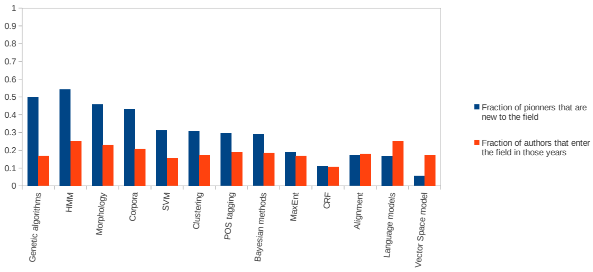

Figure 6: For each “new method”, number of “pioneers” not having published any paper before (compared to the total number of new authors during the same period of time).

For this study we only take into account authors who have published at least 5 papers in the ACL Anthology, in order to take into consideration authors who have contributed to the domain during a period of time relevant for the study. We consider as “pioneers” the authors of the first 25% of papers in which a keyword referring to a method is introduced (for example, the first papers where the keywords *support vector machine* or *SVM* appear). We then calculate, among this set of authors, the ones who can be considered as new authors, which means people who have not published before in the field. Since there are every year a large number of new authors (who use standard techniques) we compare the ratio of new authors using new techniques with the number of authors using already known techniques over the considered period. Results are visible in figure 6.

Results are variable depending on the method under consideration but some of them seem interesting. Papers with the keyword Hidden Markov Model in the 1990s seem to be largely written by new comers, probably by researchers having tested this method in related fields before (and we know that it was the case of Jelinek's team who was largely involved in speech processing, a domain not so well represented in the ACL Anthology before the 1990s. Of course, Jelinek and colleague were confirmed and even highly established researchers already at the beginning of the 1990s). We observe a similar patten for genetic algorithms but the number of authors is too limited to say if the trend is really meaningful. SVM also seem to have been popularized by new comers but it is not the case of language models or of the vector space model. A more thorough study is of course needed to confirm and better understand

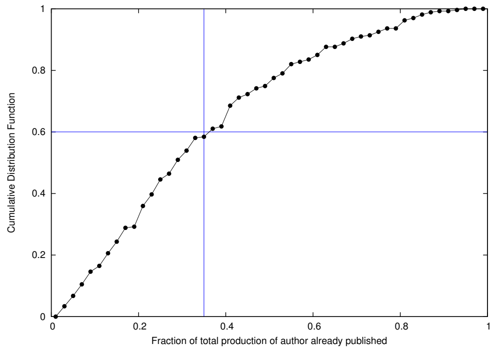

Figure 7: Distribution function of the number of papers already published by "pioneers" when they have published their paper on the new method, compared to the total production of their career.

these results.

We then do a similar experiment to try to determine when, during their career, researchers use new methods. Practically, we examine at what point of their career the authors who are characterized as "pioneers" in our study (what refers to the first authors using a new method) have published the papers containing new methods (for example, if an author is one of the first who employed the keyword SVM, has he done this at the beginning of his career or later on?). The result is visible in figure 7 and shows that 60% of pioneers had published less than a third of their scientific production when they use the new method. We thus observe a similar set of authors between the pioneers and researchers having published so far in related but nevertheless different communities. To confirm this result, it would be useful to study other domains and other corpora (in computer science, linguistics, cognitive sciences) so as to get a better picture of the domain, but the task is then highly challenging.

One may want then to observe the diversity of methods employed in the domain, especially by the set of people called "pioneers" in our study. Figure 8 shows in blue the number of methods detected for the pioneers and in red the number of methods used by all the authors.

We see that pioneers, when taking into consideration the whole set of papers in the ACL Anthology, are using a larger number of methods. They are over represented among authors using 3 methods and more. This group of people also contribute to a larger number of sub-areas in the domains compared to the set of other authors.

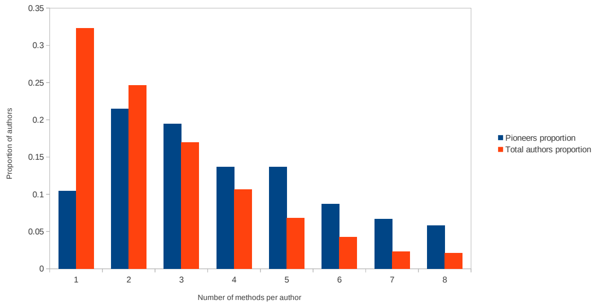

Figure 8: Proportion of "pioneers" experts in a given number of methods compared to all the other authors in the corpus.

4 Conclusion

We have presented in this paper an analysis of the ACL Anthology corpus. Our analysis is based on the identification of keywords which are categorized according to their informational status. Categorization is done according to a Text Zoning analysis of the papers' abstracts, which provides very relevant information for the study. We have shown that coupling keyword extraction with Text Zoning makes it possible to observe fine grained facts in the dynamics of a scientific domain.

These tools only give pieces of information that should be confirmed by subsequent studies. It is necessary to go back to the texts themselves, consult domain experts and probably the larger context to be able to get a really accurate picture of the evolution of a scientific domain. This multi-disciplinary research means that to collaborate with people from other fields is needed, especially with the history of science and epistemology. However, the platforms and the techniques we have described in this paper are now available and can be re-used for other kinds of studies, making it possible to reproduce similar experiments across different domains.

References

Ashton Anderson, Dan Jurafsky, and Daniel A. McFarland. 2012. Towards a computational history of the acl: 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 13–21, Jeju Island, Core. Association for Computational Linguistics.

Didier Bourigault and Christian Jacquemin. 1999. Term extraction + term clustering: An integrated platform for computer-aided terminology. In *Proceedings of the Ninth Conference on European*

Chapter of the Association for Computational Linguistics, EACL '99, pages 15–22.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Michel Callon, John Law, and Arie Rip. 1986. *Mapping the dynamics of science and technology*. McMillan, London.

Michel Callon, Jean-Pierre Courtial, and Françoise Laville. 1991. Co-word analysis as a tool for describing the network of interaction between basic and technological research: The case of polymer chemistry. *Scientometrics*, 22(1):155–205.

Katarina Frantzi and Sophia Ananiadou. 2000. Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Eugene Garfield. 1972. Citation Analysis as a Tool in Journal Evaluation. *Science*, 178(4060):471–479.

Michelle Girvan and Mark E J Newman. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99:7821–7826.

Roger Guimera, Brian Uzzi, Jarrett Spiro, and Luis A. Nunes Amaral. 2005. Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science*, 308(5722):697–702.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. 2011. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 273–283, Edinburgh.

Yufan Guo, Roi Reichart, and Anna Korhonen. 2013. Improved information structure analysis of scientific documents through discourse and lexical constraints. In *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, pages 928–937.

James Curran and Stephen Clark and Johan Bos. 2007. Linguistically Motivated Large-Scale NLP with C&C and Boxer. In *Proceedings of the 45th Meeting of the Association for Computation Linguistics (ACL)*, pages 33–36.

Gary Geunbae Lee, Jong-Hyeok Lee, and Jeongwon Cha. 2002. Syllable-pattern-based unknown-morpheme segmentation and estimation for hybrid part-of-speech tagging of korean. *Computational Linguistics*, 28(1):53–70.

Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics*, 75(6):468–487.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 2007. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 3 edition.

Roi Reichart and Anna Korhonen. 2012. Document and corpus level inference for unsupervised and transductive learning of information structure of scientific documents. In *Proceedings of COLING (Posters)*, pages 995–1006, Mumbai.

Anna Ritchie, Stephen Robertson, and Simone Teufel. 2008. Comparing citation contexts for information retrieval. In *Proeedings of the 17th Conference on Information and Knowledge Management (CIKM)*, pages 213–222, Napa Valley.

Henry G Small. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of American Society for Information Science*, 24(4):265–269.

Imad Tbahriti, Christine Chichester, Frédérique Lisacek, and Patrick Ruch. 2006. Using argumentation to retrieve articles with similar citations: An inquiry into improving related articles search in the medline digital library. *I. J. Medical Informatics*, 75(6):488–495.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Articles*. University of Edinburgh.