Automatic Annotation of Semantic Term Types in the Complete ACL Anthology Reference Corpus

Anne-Kathrin Schumann, Héctor Martínez Alonso
ProTechnology GmbH, Dresden, Germany, Thomson Reuters Labs, Toronto, Canada
annek_schumann@gmx.de, hector.martinezalonso@thomsonreuters.com

Abstract

In the present paper, we present an automated tagging approach aimed at enhancing a well-known resource, the ACL Anthology Reference Corpus, with semantic class labels for more than 20,000 technical terms that are relevant to the domain of computational linguistics. We use state-of-the-art classification techniques to assign semantic class labels to technical terms extracted from several reference term lists. We also sketch a set of research questions and approaches directed towards the integrated analysis of scientific corpora. To this end, we query the data set resulting from our annotation effort on both the term and the semantic class level level.

Keywords: semantic labeling, terminology, history of science

1. Introduction

Science changes continually: While certain research topics may be in a state of stagnation or decline, other research fronts move forward rapidly. However, even “dormant” (Menard, 1971) science can regain importance if new data is produced or methods are developed to tackle unresolved research problems. Scientific thought exhibits intricate evolutionary patterns (Fleck, 1980) and paradigm change (Kuhn, 1962) can affect the structure and outline of a whole discipline.

The availability of large collections of digitized scientific text enables systematic studies of the processes that drive scientific development. Recent years have seen a notable increase in quantitative studies of scientific text collections, e. g. Hall et al. (2008), Gupta and Manning (2011), Michaelis et al. (2013), Mariani et al. (2014), Babko-Malaya et al. (2015), Schumann and QasemiZadeh (2015b), Asooja et al. (2016), Francopoulo et al. (2016), Schumann (2016), Heyer et al. (2016).

The present study is a contribution to this research strand. Our work centers on the use of semantic labeling techniques for the automatic enhancement of a small corpus of manual term and semantic class annotations. The ultimate goal of our work, however, is to use this information as one feature in the profiling of scientific papers, communities, and disciplines. In using semantic class labels as one source of information, we take a macro- rather than a micro-perspective: While individual words and terms are certainly indicators of scientific trends and developments, it is necessary to relate them to more coarse-grained categories for an overall view of a scientific discipline. Semantic class annotations allow us to answer detailed questions about the evolution of computational linguistics over time. The data set described in this paper is made available to the research community[1].

2. Motivation and Related Work

The present investigation is conceptually related to earlier studies dedicated to the lexical analysis of diachronic corpora. Since the well-known work by Hall et al. (2008), topic modeling has been widely employed in the analysis of diachronic data. Topic modeling, however, has the disadvantage that it is ignorant to the concept of domain relevance. Topics, therefore, have to be painstakingly inferred post-hoc from word sets, and it is not straightforward which conclusions can be made on the basis of a topic model. Later work has shown that interesting insights can be obtained even with relatively simple methods of analysis, if the domain terminology is used as a clean, high-quality lexical representation of the data (Schumann and QasemiZadeh, 2015b; Schumann, 2016; Heyer et al., 2016). The present study continues this line of research by relating individual terminological units to coarse-grained semantic classes. In particular, by adding semantic class information to existing knowledge about terminological units, we enable multi-dimensional queries of the data. On the basis of terminological and semantic class information, we can, for example, ask not only which terms have been trending in computational linguistics at a given time, but we can study the evolution of various sub-fields of computational linguistics and check which associations individual terms form within these sub-fields. This does, however, not prevent us from “drilling down” to the level of individual terms, but, if necessary, we can also take a more coarse-grained perspective by “zooming out” from there, as in traditional OLAP-style[2] analyses. In sum, we believe that our approach provides two types of added value if compared to earlier research:

- The lexical basis of our investigation has a sound terminological foundation. The lexical items analyzed are not random words, but they have been identified as relevant by subject-matter experts.
- We add a second level of analysis and thus provide the means for more expressive queries of the data.

Technically, our study is related to well-established lines of work in taxonomy enrichment and domain knowledge-base population (Montoyo et al., 2001; Bergamaschi et al., 2007;

[1] `https://github.com/anetschka/complingterm.`

[2] Online analytical processing (Codd et al., 1993).

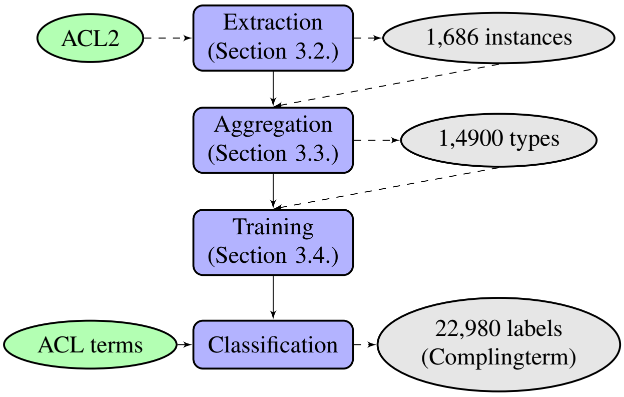

ACL2
Extraction (Section 3.2.)
1,686 instances
Aggregation (Section 3.3.)
1,4900 types
Training (Section 3.4.)
ACL terms
Classification
22,980 labels (Complingterm)

Figure 1: Flowchart representation of the semantic annotation process.

Pekar and Staab, 2003; Ruiz-Casado et al., 2007; Popescu et al., 2008; Ji and Grishman, 2011) in that we aim at the automatic type-based estimation of the semantic class of words or word sequences. Inspired by this family of approaches, we label each term, which is a specialized nominal expression of length one or more, with a semantic class.

3. Data Preparation and Semantic Labeling

We work on the ACL Anthology Reference Corpus (ACL ARC) in its first version (Bird et al., 2008). This corpus contains more than 10,000 scholarly articles from the computational linguistics domain that were published between 1965 and 2006.

We also use two additional data sets that have been created on the basis of the ACL ARC. In particular, we use a list of technical terms (ACL RD-TEC 1.0, termed *ACL1*) that was created by means of automatic term extraction (Q. Zadeh and Handschuh, 2014). More specifically, the term list was created with the help of several term extractors, and each term candidate was then manually validated by the main curator of the resource. This process resulted in more than 20,000 specialized terms that were deemed valid.

In our experiments, the ACL1 term list is used to identify all known terms. Moreover, we use a set of in-line, double-blind term and semantic class annotations (ACL RD-TEC 2.0, termed *ACL2*) provided on a subset of abstracts from the ACL ARC (QasemiZadeh and Schumann, 2016). These annotations were created by two human annotators in a multi-step process that resulted in both term span and semantic class annotations, following annotations guidelines that differentiate between seven semantic classes, as shown in Table 1. We use these high-quality annotations to train our classification models. Figure 1 provides a graphical representation of our work-flow for data preparation and annotation. Data flows from green input data ellipses to gray output data ellipses are represented with dashed lines. Blue boxes represent major work steps in the process and are explained in the sections to follow.

3.1. Data Preparation Work-flow

Annotation instances from ACL2 can be divided into 3 categories:

- Perfect matches: identical term spans marked by both annotators
- Partial matches: overlapping, but not identical term spans
- Annotation conflicts: term spans marked by only one of the two annotators

Type	Example
Technologies	parsing
Tools	parser
Language resources	corpus
Lang. resource products	Brown corpus
Models	language model
Measures	Bleu score
Other	*residual class*

Table 1: Semantic classes in ACL2.

Each instance has at least one semantic class assigned to it, but in all categories multiple (conflicting) class assignments can occur. Table 1 shows which semantic classes have been annotated in ACL2 (see Schumann and QasemiZadeh (2015a) for details). We have prepared our training and test data as follows:

1. We extracted reliable *annotation instances* from ACL2. This is explained in more detail in Section 3.2.
2. We created consistent *annotation types* by aggregating annotation instances by their term lemmas. We relabeled a part of the annotations and merged the original semantic classes into larger containers. This procedure is explained in Section 3.3.
3. We merged both term lists (ACL1+2) to create a maximal term list and identified term occurrences in the whole ACL ARC.
4. The classifier is described in Section 3.4.

3.2. Extraction of Reliable Annotation Instances

From ACL2, we extracted all consensual annotations, that is, term occurrences that were annotated by both annotators with the same span and semantic class: Among the 4,849 manual annotation instances, 2,583 share exactly the same span, being complete matches of each other. Among those, 1,686 also share the same semantic class. Appendix A shows how these terms are distributed over the different publication years in the data set and how many abstracts were annotated for each year.

As can be seen from the table, the distribution is highly skewed in favor of an, overall, too large “other” class. Why is this data so unbalanced? In many cases, the “other” class contains linguistic units that are neither language resources nor language resource products. Examples of such “other” instances are terms like “verbal interaction” or ”Japanese kanji-kana characters”. However, linguistic left-overs do not make up the complete “other” class: Specialized language is embedded in discourse and, therefore, terminological units in real-world abstracts are sometimes not those that one might find in a specialized taxonomy or ontology. For example, they might be ambiguous if taken out of context, or they might be discoursive variants of known terms that still bear terminological weight. Examples of such items are terms like “syntactic descriptions” (which might be anything between a strongly formalized and a free-form description) or “error characters”. Such units are, indeed, characteristic of academic language and should not be ignored. However, with the heavy skew observed, the ACL2 data set seems hardly usable for automatic prediction.

3.3. Conversion of Annotation Instances to Annotation Types

Since our work was aimed at creating a reliable set of semantic term type annotations, *instance* annotations from ACL2 had to be converted to annotation *types*. We did this by grouping the 1,686 perfect matches from ACL2 by their term lemmas, arriving at 1,490 annotation types. All “other” attributions were then reconsidered and, if necessary, relabeled manually. The main goal of this step was to arrive at a more even class distribution in the training data and, most notably, a smaller residual class. Since in the ACL2 data identical lemmas can have diverging semantic labels in different annotation instances, re-annotation with the aim of producing consistency with respect to other annotation *instances* seemed justified. All relabellings were discussed by both authors of this paper. If necessary, ACL papers from the corpus were analyzed in detail. Then, to deal with very tiny classes, we merged the 7 semantic classes originally annotated in ACL2 into 4 larger classes, namely:

- **Mathematics**: This class contains the Models and Measures classes from ACL2.
- **Technologies**: This is the superclass of the ACL2 classes Technologies and Tools.
- **Linguistics**: This class contains terms with a linguistic background, that is the Language resources and Language resource products classes from ACL2 along with a relevant share of relabeled “other” instances.
- **Interdisciplinary**: After the re-annotation, what is left of the “other” class now contains general, higher-level interdisciplinary terms that nevertheless bear terminological weight.

The 4 classes were formed by merging conceptually similar tiny classes to form larger and slightly more general classes. Figure 2 provides a graphical overview of the semantic classes before and after merging. Green boxes represent the 7 semantic classes that were manually annotated in ACL2. These were merged into 4 coarse-grained classes for more reliable automatic prediction (blue boxes). The tree in Figure 2 is also labeled with example terms from ACL2[3]. As a result of our restructuring of the data, the bulk of the purely linguistic terms has been moved to the Linguistics class. What remains in Interdisciplinary can now be related to technicalities of the scientific process and to scientific discourse. Examples of the Interdisciplinary class are terms such as “speech-act indirectness”, “np-hard problem”, or “telephone communication”. Table 2 gives an overview of the training data resulting from our preparatory work. The table shows that our efforts have produced a more even class distribution. Classes with very few instances have been merged with larger classes.

Type	Number
Mathematics	226
Technologies	677
Linguistics	283
Interdisciplinary	304
Overall	**1,490**

Table 2: Term type distribution in training data

3.4. Automatic prediction

In order to assign a semantic class to the unlabeled terms extracted from ACL1, we have implemented a logistic-regression classifier trained on the annotated data (ACL2). We have performed feature selection using ten-fold cross-validation. The resulting classifier uses the following features:

1. **BoW**: Identity of the term headword. If the term is longer than one word, we treat all words from the second as a bag of words (BoW). In this way, we give the headword of the term a special treatment, which makes it easier to identify as a trigger term for a certain class.
2. **Length**: The length of the term in number of words, and the proportion of capitalized characters. These features help identify multi-word expressions and determine whether they are terminological units, or whether they are acronyms.
3. **Brown**: The Brown clusters (Brown et al., 1992) from the full ACL corpus for the words in the term. Brown clusters group words in a corpus according to their immediate surrounding bi-grams and provide good features to estimate semantic classes.
4. **Embeddings**: The average word embedding for all words in the term. We use embeddings from the ACL corpus with 100 dimensions and a word window of 5. Using embeddings allows us to incorporate distributional information of words involved in a term that is larger in scope than the information captured by Brown clusters.
5. **WordNet**: The number of senses in WordNet (Miller, 1995) for the term headword, as well as the list of semantic types (e.g. *noun.cognition*) for these senses. The intuition behind these features is that more polysemous words are more likely to be the head of terms, and by including their possible coarse-grained senses it becomes easier to characterize their semantic class.

We have used ten-fold cross-validation during development to determine which classifier setup was more robust. However, the classifier we used to actually tag the corpus is trained on the full dataset. In this way, the scores we provide are a conservative estimate of the actual performance of the classifier. Table 3 shows the performance of the classifier in terms of F1 score for the chosen classifier, which uses all the features listed, and a comparison baseline that only uses **BoW**.

As can be seen from the table, our classifier performs reasonably well, reaching an average F1 score of 0.73. Classification performance is obviously influenced by the number of training examples with the largest class achieving the

[3] Note that due to our work-flow, we have access to both coarse-grained and fine-grained (ACL2) semantic information for the 1,490 training terms. For all remaining terms we have coarse-grained semantic labels.

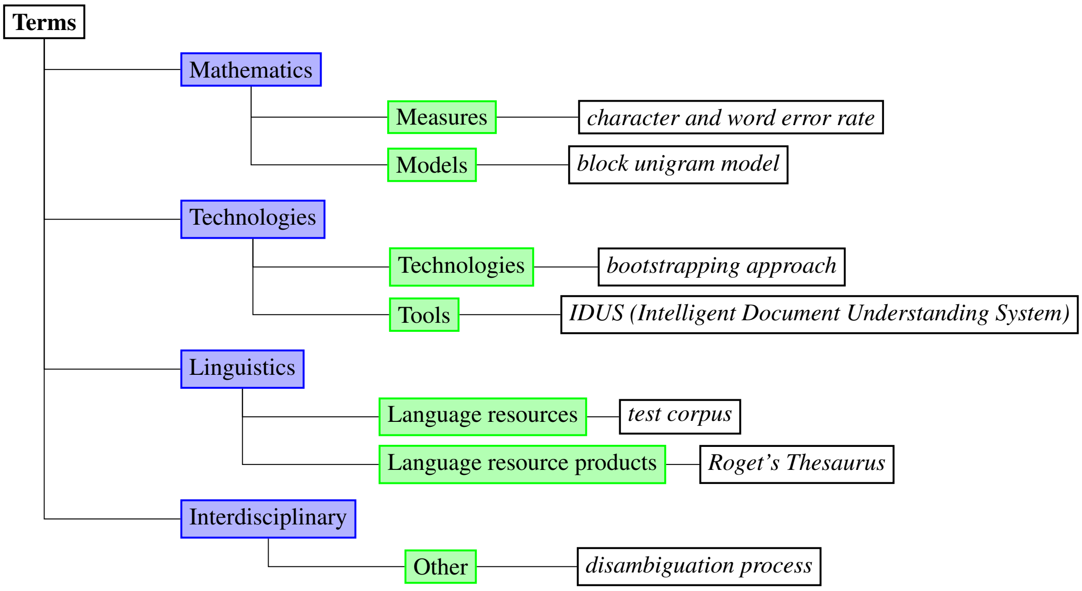


Figure 2: Taxonomic representation of the semantic categories used for training data preparation. Leaves in the tree are example terms extracted from ACL2 data.

	Baseline	**Full**
Technologies	0.75	0.83
Linguistics	0.61	0.70
Math. concepts	0.59	0.64
Interdisciplinary	0.53	0.60
Micro-Average	**0.65**	**0.73**

Table 3: F1 classification scores.

best individual score. The rather fuzzy in terms of distribution and lexically very varied Interdisciplinary class exhibits the lowest score. Conversely, the Technologies class is easier to identify because it is often made up of longer expressions and tends to show frequent, informative words like *system*.

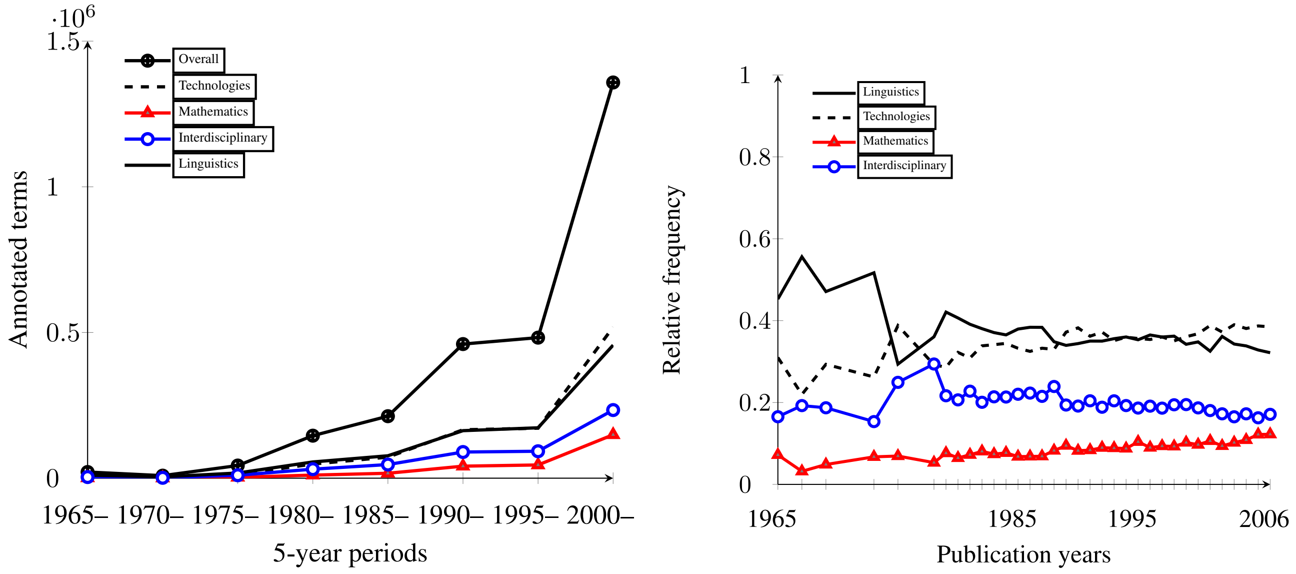
It is not straightforward to compare our classification result to the outcome of other annotation efforts, not only because there are only few such efforts, but also because annotation evaluation scenarios vary considerably. However, QasemiZadeh and Schumann (2016) provide a detailed analysis of their manual annotation of the same data set. They report class-wise IAA scores that range between 0.44 (for the Model class, compare Figure 2) and 0.83 (for the Tool class, which is easily identifiable, but too small for classification, compare Appendix A), respectively. This shows that in manual annotation, too, class parameters influence annotation quality. The overall quality of our automatic classification, therefore, seems reasonably close to that of the current manual annotation benchmark. It must be noted, however, that the score reported by QasemiZadeh and Schumann (2016) is a combined score for both term identification and semantic class assignment.

We also performed a manual analysis of a small subset of the predictions, namely a 30-term subset for each of the four classes. Out of thirty examples of the Technologies category, we only would relabel one ("bottom-up approach") as part of Interdisciplinary. In the predictions for Linguistics, we find the false positive "electronic mail", which should be a Technology. This mis-prediction is a result of the WordNet feature providing the super-sense *noun.communication* to the word "mail". The Mathematics category ends up being the label of choice for all expressions with the word "model" as a headword ("IBM Models 1-2"). "Model" is a fairly polysemous word that could yield terms of any category, but the frequency and the annotation preference for labeling them as terms in Mathematics enforces this bias. The Interdisciplinary class does indeed contain more noise, and has many terms that should belong to the Linguistics category ("simultaneous speech" or "subject-object relation"). We attribute the permeation between these two classes to some of the features that give account for polysemy, as both Linguistics and Interdisciplinary contain more words with numerous possible senses, in addition to being the classes with highest word variety in terms of type–token ratio. Our manual analysis corroborates that Technologies is by far the most reliable label of our predictions. Appendix B shows some of the examples that underwent the manual analysis just described. For each term, the table shows the classification probabilities given by our classifier for the four possible classes. The value for the resulting class is highlighted in bold.

4. Analysis of Resulting Data Set

4.1. Structure

The data set resulting from our annotation is essentially an annotated term list. This list contains 22,980 computational linguistics terms. Each term is annotated with a probability vector with four components, where each component represents the probability that a given term is an instance of the respective semantic class, as illustrated in Appendix B. Finally, a class label indicates the semantic class with the highest probability.



(a) Number of term occurrences with semantic label over time.

(b) Relative importance of semantic classes over time.

Figure 3: Classification results: annotated terms and semantic classes over time.

4.2. Querying the Data

As pointed out earlier, we maintain that semantic class annotations facilitate the analysis of scientific text corpora. In the last section of this paper, we, therefore, sketch some of the research questions that can be tackled through multidimensional analyses involving both the term and the semantic class level.

One of the most prominent properties of the ACL data as a whole is its exponential growth over time. This finding is supported by Figure 3a which plots the number of known and semantically labeled terms over time periods of 5 years. The tendency of scientific disciplines to grow exponentially has been described already by Menard (1971) and it is, indeed, a property that informs all studies on the ACL data. Figure 3b plots the classification results over time, when class-wise term counts are normalized by the number of all terms per year. Thus, the plot shows the relative importance of each semantic class over the course of the publication period. The probably most prominent trend in this plot is that Linguistics and Technologies, over the years, switch ranks, with Linguistics slightly losing importance and Technology mentions becoming more frequent. We also observe a slight, but steady increase in the relative frequency of terms referring to mathematical concepts.

To check whether our dataset allows us to ask more specific questions about the development of computational linguistics, we set up a simple database holding information about both terms, their semantic classes and term occurrences across all publication years. We have used this database to extract frequency data that was further analyzed with the following techniques:

- We used word rank comparisons to identify "trending" terms, that is, lexical units that gain importance in a given period, as proposed by Schumann and Qasemi-Zadeh (2015b). The method is based on the comparison of two ranked word lists for consecutive time periods and identifies lexical units that undergo a strong positive rank shift, that is, words that exhibit an "upwards" movement in the transition from one time period to the next.

1986–1989	2000–2006
tree	measure
formalism	language model
user model	f-measure
representations	n-gram
domain model	distribution
discourse model	f-score
measure	n-grams
perplexity	nist
language model	translation model
projection	parse tree

Table 4: Mathematics terms with high positive frequency rank shifts in two different time intervals. In both cases, frequency counts were compared to frequency counts from the preceding time interval.

- We used frequency and productivity scores to analyze individual terms' life cycles, as proposed by Schumann (2016). In this approach, productivity is formalized in terms of entropy, that is, a base term with many and frequently used related multi-word units is considered particularly productive. Unlike simple frequency analysis, this approach helps, for instance, to differentiate between short-term tendencies and longer-term trends.

In a first step, we used our SQL database to obtain an overview of the terms pertaining to the four semantic classes and their frequencies at various points in time. For

distribution		**measure**	
1986–1989	1990–1996	1986–1989	1990–1996
distribution	distribution	measure	measure
probability distribution	probability distribution	distance measure	f-measure
gaussian distribution	uniform distribution	similarity measure	similarity measure
binomial distribution	joint distribution	evaluation measure	evaluation measure
class distribution	normal distribution	association measure	distance measure
probability distribution matrix	frequency distribution	confidence measure	statistical measure
	binomial distribution	theoretic measure	f measure
	distributional similarity		cosine measure
	prior distribution		precision measure
	gaussian distribution		association measure
	...		...

Table 5: Top terms matching “distribution” and “measure” in two time periods. Terms are sorted by their frequencies.

instance, we counted the occurrences of terms pertaining to Mathematics in 4 time intervals: 1980-1985, 1986-1989, 1996-1999, and 2000-2006. Using the rank comparison technique, we could then contrast different methodologies and models that were popular in computational linguistics research at different times. Table 4 exemplifies this by giving an overview of the terms with the highest positive frequency rank shifts from Mathematics for two different time intervals.

On the basis of the comparison exemplified in Table 4, it seems relatively easy to contrast explicit, knowledge-based or abstract modeling in the 1980s with scoring and statistical analysis, approaches that were prominently used in the later time period. However, the table also seems to indicate that there is nothing disruptive about this change in the Mathematics class. Rather, concepts like “measure” or “language model” seem to have kept gaining importance, whereas other concepts seem to have stagnated. This claim can easily be substantiated by querying our database for frequencies and collocations containing terms such as “feature”, “distribution”, “measure”, or “model”. These terms exhibit strong productivity increases, that is, a growing number of related multi-word units, in the 1980s and a continuous increase in frequency over the complete period of time under study. This means that these terms, at the time of their rise, did not necessarily denote completely novel concepts, but they have found new fields of application or new ways were found to build and apply such models, features, etc. In this sense, terms like “distribution” or “measure” can be hypothesized to constitute motors of scientific innovation. Table 5 details the results of example queries for the terms “distribution” and “measure” for only two time periods, namely 1986–1989 and 1990–1995. The rise in productivity for the two terms over this short time period seems rather strong, however, as claimed before, the examples illustrate a scientific evolution rather than a disruptive change. It is an interesting task to identify the area of computational linguistics that has initiated this development, and with the rich data set presented here, it seems actually feasible to solve this task.

By repeating the analysis for the Technologies class we were able to highlight some interesting technological trends that have been prominent in computational linguistics over the last decades. For the 1990–1995 interval, for instance, the ranks shift analysis highlights the following trends: the use of WordNet as a lexical resource, the use of finite-state transducers for complex analysis tasks, work on word sense disambiguation, but also on part-of-speech tagging and other kinds of linguistic annotation, to mention only the most prominent items in our result list. Techniques used are not purely statistical, as highlighted by terms such as “lexical rule” or “heuristic”.

For the 2000–2006 period, many of the lexical units highlighted by the rank shifts analysis in Technologies are related to recent work on the use of machine learning in computational linguistics, e. g. terms such as “classifier” or “feature”. However, ontologies have also gained importance and the 2000–2006 slice of Technologies includes novel lexical units such as “ontology learning”, “ontology acquisition”, or “ontology induction”. Annotation, both manual and automatic, remains an important topic and FrameNet is introduced as a new resource. The growing use of the term “NLP” and the rise of statistical machine translation seem to constitute other important tendencies.

5. Conclusion

In this paper, we have described our methodology for creating a large data set of semantically annotated terms. We have merged linguistic information from several existing data sets and performed manual re-annotation to arrive at a more even distribution of semantic class instances. We have used state-of-the-art classification techniques to provide semantic class labels for known term instances. Both manual and automatic evaluation of our classification results indicate reasonably good classification quality.

We have also carried out an initial analysis of the resulting data set to exemplify which kinds of questions can be answered using the rich annotations provided by our data set. Although we have by no means described a complete methodology for the analysis of scientific corpora, we believe that our analysis shows what is possible with the wealth of data available. State-of-the-art methods can be used not only to identify large-scale trends such as the growing importance of statistical and machine learning methods in computational linguistics. The semantic class labels resulting from our classification effort allow to per-form this analysis on a more fine-grained level and search for interesting phenomena in actual subfields of computational linguistics. We have also shown that the analysis of mere frequency and productivity information allows us to identify some of the concepts that have in the past served as motors of innovation, such as the concepts "measure" and "distribution". In the future we hope to develop an integrated methodology that will be able to reliably relate terms and groups of terms to specific time periods and subfields and to trace scientific innovations back to the nuclei from which they originated.

6. Bibliographical References

Asooja, K., Bordea, G., Vulcu, G., and Buitelaar, P. (2016). Forecasting Emerging Trends from Scientific Literature. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association.

Babko-Malaya, O., Seidel, A., Hunter, D., HandUber, J. C., Torrelli, M., and Barlos, F. (2015). Forecasting Technology Emergence from Metadata and Language of Scientific Publications and Patents. In *Proceedings of the 15th International Conference on Scientometrics and Informetrics (ISSI 2015)*. Boğaziçi Universitesi.

Bergamaschi, S., Bouquet, P., Giazomuzzi, D., Guerra, F., Po, L., and Vincini, M. (2007). An Incremental Method for the Lexical Annotation of Domain Ontologies. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 3(3):57–80.

Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., Lee, D., Powley, B., Radev, D., and Tan, Y. F. (2008). The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. European Language Resources Association.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479.

Codd, E., Codd, S. B., and Salley, C. T. (1993). Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandat. Technical report, E.F. Codd & Associates.

Fleck, L. (1980). *Entstehung und Entwicklung einer wissenschaftlichen Tatsache. Einführung in die Lehre vom Denkstil und Denkkollektiv*. Suhrkamp, Frankfurt am Main. First edition in 1935, Benno Schwabe & Co., Basel.

Francopoulo, G., Mariani, J., and Paroubek, P. (2016). Predictive Modeling: Guessing the NLP Terms of Tomorrow. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association.

Gupta, S. and Manning, C. D. (2011). Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*. Asian Federation of Natural Language Processing.

Hall, D., Jurafsky, D., and Manning, C. D. (2008). Studying the History of Ideas Using Topic Models. In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Heyer, G., Kantner, C., Niekler, A., Overbeck, M., and Wiedemann, G. (2016). Modeling the dynamics of domain specific terminology in diachronic corpora. In *Proceedings of 12th International conference on Terminology and Knowledge Engineering (TKE 2016)*.

Ji, H. and Grishman, R. (2011). Knowledge Base Population: Successful Approaches and Challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*. Association for Computational Linguistics.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Mariani, J., Paroubek, P., Francopoulo, G., and Hamon, O. (2014). Rediscovering 15 Years of Discoveries in Language Resources and Evaluation: The LREC Anthology Analysis. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. European Language Resources Association.

Menard, H. W. (1971). *Science: Growth and Change*. Harvard University Press, Cambridge, Mass.

Michaelis, J. R., Mcguiness, D. L., Chang, C., and Babko-Malaya, O. (2013). Towards explanation of scientific and technological emergence. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2013*.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Montoyo, A., Palomar, M., and Rigau, G. (2001). WordNet Enrichment with Classification Systems. In *Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customisations Workshop. The Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*. Association for Computational Linguistics.

Pekar, V. and Staab, S. (2003). Word classification based on combined measures of distributional and semantic similarity. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*. Association for Computational Linguistics.

Popescu, A., Grefenstette, G., and Moëllic, P. A. (2008). Gazetiki: Automatic Creation of a Geographical Gazetteer. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital libraries*. ACM.

Q. Zadeh, B. and Handschuh, S. (2014). The ACL RD-TEC: A dataset for Benchmarking Terminology Extraction and Classification in Computational Linguistics. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*. Association for Computational Linguistics and Dublin City University.

QasemiZadeh, B. and Schumann, A.-K. (2016). The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods. In

Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA).

Ruiz-Casado, M., Alfonseca, E., and Castells, P. (2007). Automatising the learning of lexical patterns: An application to the enrichment of WordNet by extracting semantic relationships from Wikipedia. *Data & Knowledge Engineering*, 61(3):484–499.

Schumann, A.-K. and QasemiZadeh, B. (2015a). The ACL RD-TEC Annotation Guideline: A Reference Dataset for the Evaluation of Automatic Term Recognition and Classification. Technical report.

Schumann, A.-K. and QasemiZadeh, B. (2015b). Tracing Research Paradigm Change Using Terminological Methods: A Pilot Study on “Machine Translation” in the ACL Anthology Reference Corpus. In *Proceedings of 11th International Conference on Terminology and Artificial Intelligence*. CEUR Workshop Proceedings.

Schumann, A.-K. (2016). Brave New World: Uncovering Topical Dynamics in the ACL Anthology Reference Corpus using Term Life Cycle Information. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*.

Year	Lang. resources	Lang. resource products	Measures	Models	Other	Technologies	Tools	Sum	Abstracts
1978	0	0	0	0	36	4	0	40	3
1980	0	0	0	1	12	9	2	24	5
1982	0	0	0	0	16	11	4	31	3
1984	0	0	0	2	37	17	1	57	5
1986	0	0	0	1	91	10	2	104	7
1988	6	0	0	4	40	23	3	76	12
1990	14	1	2	0	52	25	0	94	11
1992	8	9	4	2	75	60	6	164	19
1994	5	1	2	6	88	57	4	163	16
2001	4	0	18	14	89	88	4	217	18
2003	26	0	18	29	158	111	4	346	29
2005	17	0	20	12	89	64	2	204	23
2006	7	1	11	2	95	44	6	166	20
Overall	**87**	**12**	**75**	**73**	**878**	**523**	**38**	**1,686**	**171**

Appendix A: Time distribution of semantic classes in ACL2 (perfectly matching instances).

Term	Technologies	Linguistics	Interdisciplinary	Mathematics
general-to-specific learning	**0.97**	0.00	0.03	0.00
spoken document categorization	**0.74**	0.25	0.01	0.00
arabic stemmer	**0.89**	0.05	0.06	0.00
constraint propagation algorithm	**0.97**	0.00	0.00	0.03
token processing	**0.98**	0.00	0.01	0.01
standard grammar textbook	0.40	**0.54**	0.03	0.02
pre-discourse meaning	0.02	**0.88**	0.09	0.01
korean language	0.01	**0.96**	0.02	0.01
case particle	0.13	**0.78**	0.03	0.05
prosody	0.08	**0.52**	0.36	0.04
knowledge editing	0.30	0.00	**0.69**	0.00
multilingual information exchange	0.31	0.00	**0.68**	0.01
simultaneous speech	0.18	0.12	**0.69**	0.01
subject-object relation	0.01	0.28	**0.67**	0.05
categorisation research	0.46	0.03	**0.50**	0.00
homomorphism	0.08	0.17	0.31	**0.44**
IBM Models 1-2	0.09	0.00	0.02	**0.88**
time complexity	0.00	0.00	0.39	**0.61**
bigram distribution modelling	0.27	0.00	0.00	**0.73**
likelihood reestimation	0.43	0.00	0.00	**0.57**

Appendix B: Example output of the classifier used for manual analysis, resulting class is highlighted in bold.