limited.
Case Number 14-4079

State of the Research in Human Language Technology

A study of ACL and NAACL publications from 2007 through 2014

Karine Megerdoomian
MITRE Corporation
June 2014

Document number: MTR140208

Table of Contents

List of Tables

List of Figures

1. Purpose of Study

The goal of this study is to identify state-of-the-art in Human Language Technology (HLT) for the application areas of knowledge discovery, triage, language professional support and foundational technology, pinpointing potential research directions. We took a holistic view of the application areas to explore how the HLT community will progress in each area in the near future. We performed our analysis by examining conference publications for recent ACL (Association for Computational Linguistics) and NAACL (North American Association for Computational Linguistics) conferences, identifying terms and topics which reflect the state of HLT research.

2. Data Collection and Analysis

We collected data from the ACL Anthology site for ACL and NAACL conferences, ranging from 2007 through 2014. Data include titles, authors and year of publication for the main articles (including student papers, but not demonstration reports). In addition, we downloaded the complete articles separately[1]. Table 1 and Figure 1 show the distribution of papers.

Conference	Number of Publications
ACL-2007	204
ACL-2008	214
ACL-2009	243
ACL-2010	269
ACL-2011	343
ACL-2012	222
ACL-2013	394
ACL-2014	158
NAACL-2007	157
NAACL-2009	176
NAACL-2010	177
NAACL-2012	126
NAACL-2013	168
total	**2851**

Table 1 - Number of ACL and NAACL publications per year, 2007-2014

[1] Although some of the analyses included the abstracts, most were run on the titles of publications only. Analysis on full content is left for future research.

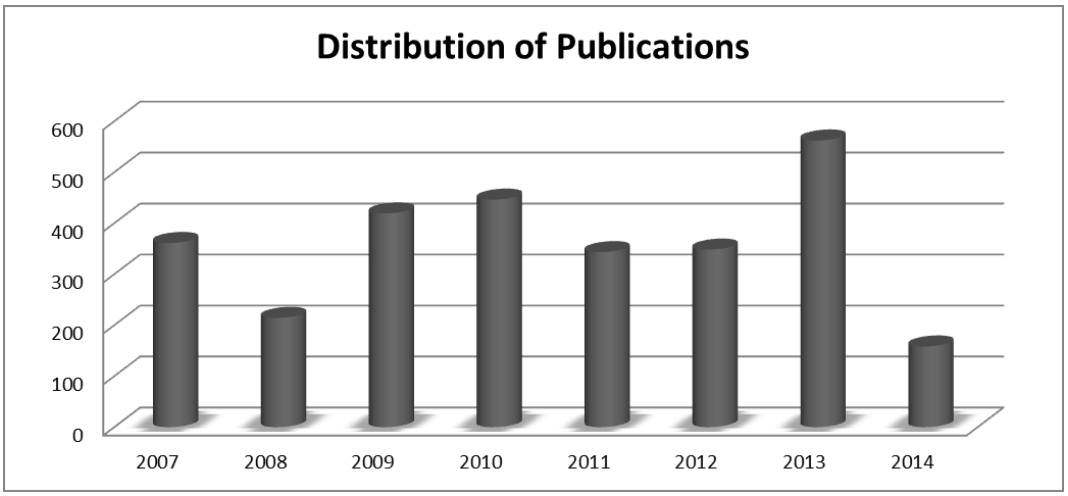

Figure 1 - Distribution of ACL and NAACL publications combined per year, 2007-2014.

Note that there are no separate NAACL publications for 2008, 2011 and 2014 because NAACL was held jointly with ACL in those years. Analysis of ACL 2014 publications only included title and author information and a pre-assigned category from ACL, as the proceedings were not yet published at the time we prepared this report.

We tagged each paper for subject areas and application category. The four application categories, or domains, are defined as follows:

- **Foundational Technology**: This domain includes enabling component technologies, data annotation and management, text-to-speech and machine translation. In addition, we also added discourse analysis and exploration of new general theories and approaches in NLP.
- **Knowledge Discovery**: The recognition and extraction of knowledge such as entities, relations, events and concepts are part of this application domain. Other areas in this category include sentiment analysis, summarization and information retrieval.
- **Triage**: This domain includes any technology that is applied for the purpose of finding documents of interest and transforming them a more computable form. These range from language and speaker/author identification to optical character recognition to topic detection.
- **Language Professional Support:** Tools that analysts and linguists can use in their daily activities, such as translation memory or computer-aided translation systems.

We provide the full list of subject areas, based on community usage, and how they fit within the top four application categories in Appendix 1. For the purposes of this study, we treat any paper that specifically mentions a subject area, as well as technology applied to that area, as belonging to the same category. Thus, papers dealing with Machine Translation (MT) evaluation and Word Sense Disambiguation specifically designed for MT are all tagged as Machine Translation and fall within the application area of Foundational Technology. For the rest of this report, *subject area* refers to the subcategories listed in Appendix 1 (e.g., Machine Translation, Sentiment Analysis, or Topic Detection) and *domain* refers to the main four application categories described above.

To capture the state of HLT research, we performed four analyses:

1. Identify the main application area or domain for each paper, studying the change in focus through the years.
2. Isolate new and emerging themes in more recent years by identifying and contrasting the main terms in the publications for each year.
3. Apply clustering to the publication date to detect the main subject areas and their relations.
4. Explore the authorship network to identify the central authors, institutions and main themes in the field.

3. Results Summary

The results show that the largest application domain for each year is the Foundational Technology category, and the most dominant subject areas are *machine translation* and *enabling component technologies*. All analyses suggest that in the field as seen through ACL and NAACL publications, statistical machine translation (SMT) dominates, and it even fuels work in other subject areas, such as *phrase-based parsing, language modeling,* and increases in *learning theory* and *corpus development.* The second most consistently published subject area in this domain is work on parsing and in particular, *dependency parsing*. Focus on low-resourced languages, such as Arabic, Chinese, Japanese and Hindi drives the work in enabling technologies such as *word segmentation* and *morphological analysis*. Another trend is the use of *discourse* analytics, especially in dialogue systems and Automatic Speech Recognition (ASR) applications.

The second prominent application domain is Knowledge Discovery, which includes research in *event and relation extraction, concept extraction* and more recently, *sentiment analysis.*

Topic Detection, which includes topic modeling, detection and classification, holds the most publications within the Triage application domain, while Language Professional Support has the least publications throughout the academic community. Of the latter, *error correction* research seems to be of interest to the research community.

Certain terms seem to be emerging or on the rise within the field. These include the use of *crowdsourcing* and analysis of *social media* (especially Twitter). More recently, *neural networks* appear frequently. It is interesting to note that *semantics,* and in particular, semantic enhancements to language models, are prominent in the field. There seems to be increasing interest in other approaches to semantics, such as *distributional semantics.* The increase of terms like *neural networks* and *distributional semantics* presumably signals the interest in deep learning for HLT applications. Overall, SMT seems to fuel the field and associated research trends, yet there is growing interest in applying knowledge-based approaches that integrate some semantic information or concept and relations understanding within the systems.

The co-authorship study demonstrates that the same few people and institutions (often focused on SMT

research and typically centered in the East Coast of the US and in China) have dominated the field for the last 8 years. Topic classification of communities also shows important research focused on semantics and knowledge extraction, while some peripheral groups (often not in the US) tend to focus on knowledge discovery research and lexical semantic approaches. The analysis of the collaboration network shows that the ACL/NAACL community exhibits small world network characteristics, with interconnected groups of authors and a few authors with important central roles.

4. Application Categories

We manually annotated the publications from NAACL-2013, ACL-2013 and ACL-2014 for subject area and domain[2]. We then used these tags to train a classifier to label the publications for the rest of the data set (i.e., NAACL/ACL papers from 2007 through 2012). After experimenting with several classifiers on the Weka package[3], we used a filtered classifier utilizing the NaiveBayesMultinomial module with default parameter settings and filtered through the StringToWordVector filter trained on n-gram frequency. We performed experiments using 10-fold cross-validation. The classifier scored 72.4% on the dataset (97.5% on the training data). The weighted f-measure was 0.691, and f-measures for the different classes are: foundationalTechnology: 0.804, knowledgeDiscovery: 0.665, triage: 0.107, languageSupport: 0.211.

The top weighted terms per domain category, extracted from the titles of ACL and NAACL publications from 2013 and 2014, are shown in Table 2 in order to illustrate the most frequent themes in each domain. In addition, Appendix 2 lists the top weighted 30 terms for each of the four application domains as obtained from the trained classifier.

Application Domain	Keywords
Foundational Technology	machine translation; statistical machine; word segmentation; semi supervised; part of speech; cross lingual; neural network; neural networks; beam search; coreference resolution; dependency parsing; POS tagging; shift reduce; latent variable; recurrent neural; large scale; dialectal arabic; chinese word; word alignment; phrase based
Knowledge Extraction	relation extraction; question answering; document summarization; distant supervision; social media; sentiment analysis; cross lingual; multi document; random walk; textual entailment; sentence compression; distributional semantics; supervision relation; matrix factorization; fine grained; surface realisation; analysis probabilistic; named entity; dual decomposition; empirical study
Triage	topic models; social media; text classification; semi supervised; weakly supervised; latent topics; topic modeling; training data
Language Professional Support	correction using; error correction; language acquisition

Table 2 - Application domains and their top keywords, based on 2013-2014 publication titles

[2] We accomplished this by studying the full papers for the 2013 publications, and the titles and their associated ACL-tagged keywords for the 2014 publications.

[3] http://www.cs.waikato.ac.nz/ml/weka/

The domain distributions for each year are provided in Figure 2 and Figure 3[4]. As the results show, the largest application domain in each year is the Foundational Technology category, followed by Knowledge Discovery. Language Professional Support has the least number of publications throughout.

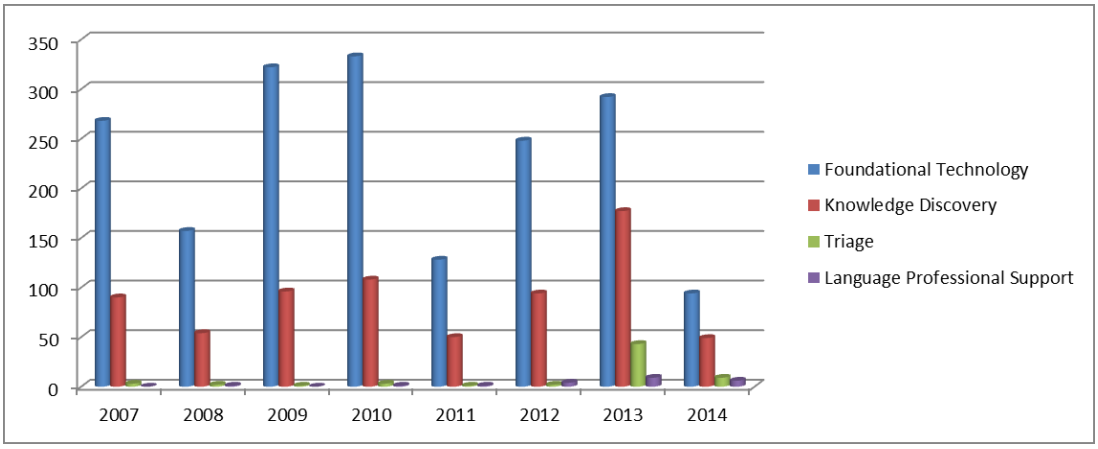

Figure 2 - Distribution of domain count per year

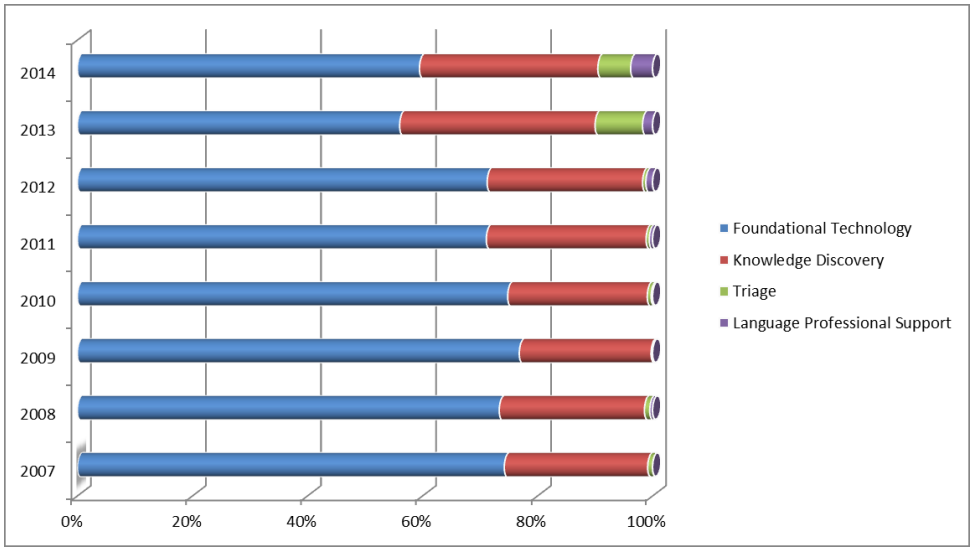

Figure 3 – Relative percentage of domains per year

For years 2013 and 2014, we manually annotated the papers for subject area category as well, and we show the distribution results in Figure 4. We provide the relative distribution of the subject areas per domain in Figure 5. As can be seen from these charts, *enabling component technologies* and *machine translation* from the Foundational Technology application domain are the largest categories overall. The other set of important subject areas are from the Knowledge Discovery domain and include *relations, event,* and *concept extraction* as well as *sentiment analysis.*

[4] The results of the classifier seem to be slightly skewed towards the first two classes and may underrepresent the number of articles on triage or language support.

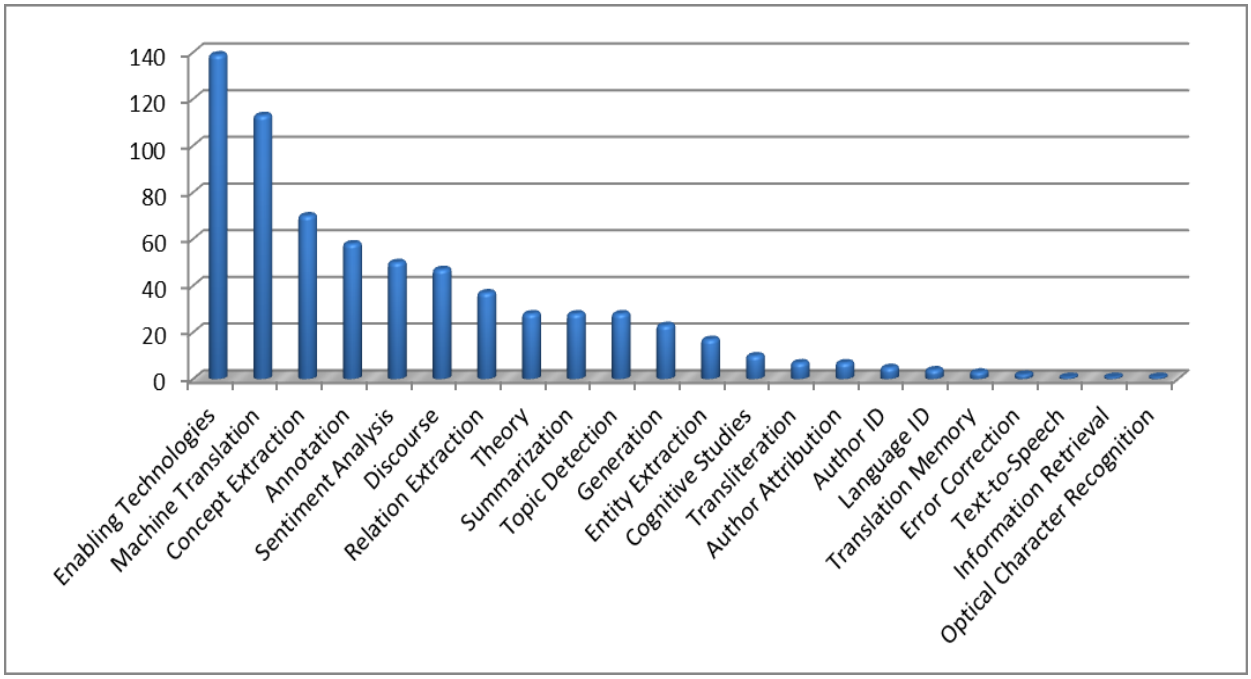

Figure 4 - Distribution of subject area counts for 2013 and 2014 publications

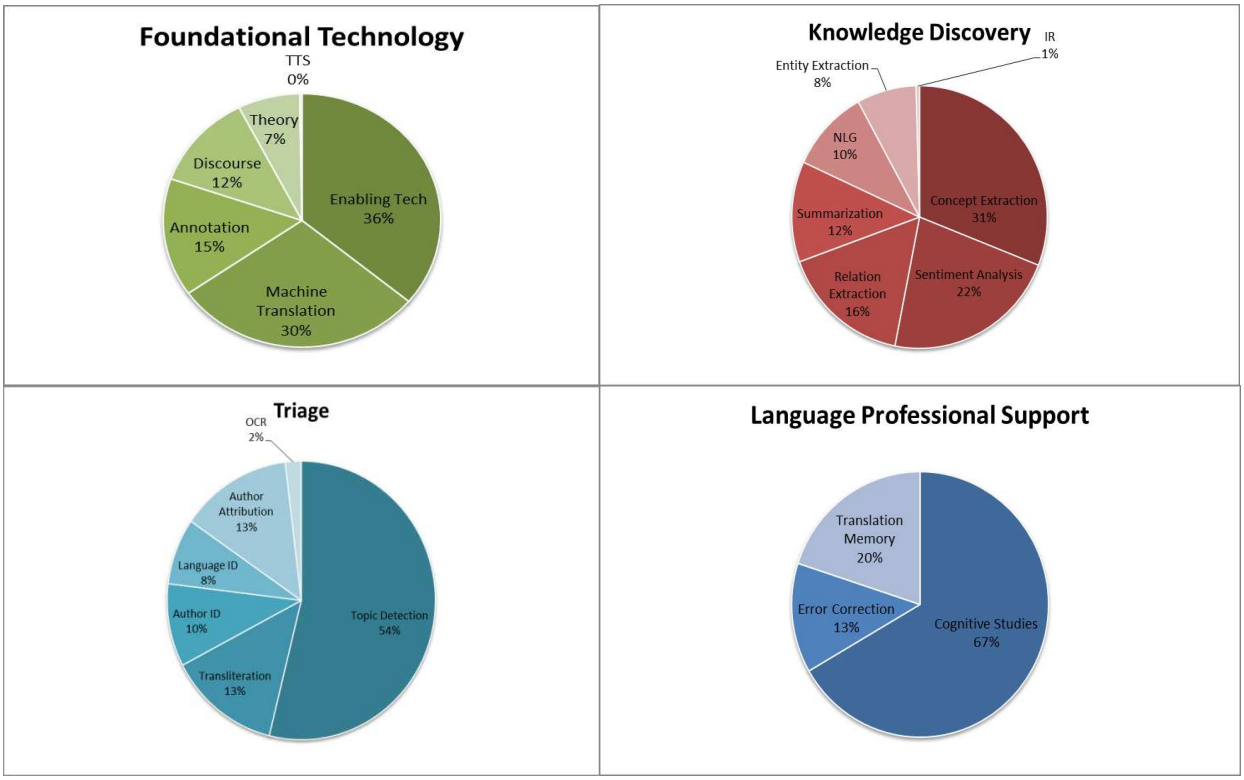

Figure 5 - Relative distribution of topics per application domain for 2013 & 2014 papers

5. Topic Detection

Main themes

To identify the main topics across the years, we extracted the most frequent terms for each year from the titles and abstracts of the publications. First, we ranked the top 20 bigrams per year based on their likelihood frequency (see Appendix 3 for the list of terms[5]). Of these, terms that represented language reflecting abstract writing style were ignored; including terms such as *state of the art, significant improvements, paper proposes,* or *experiments show*. The remaining terms are therefore substantive terms that represent a research focus in the field of HLT and appear in the top ranks across the years. Table 3 lists these terms for each year, in order of frequency of occurrence.

Year	Keywords
2007	machine translation; statistical machine; coreference resolution; dependency parsing; named entity; information retrieval; question answering; semi-supervised; sense disambiguation; support vector; vector machines; domain adaptation; natural language; mandarin broadcast; phoneme conversion; word sense; language processing; speech recognition; semantic relatedness; semantic role
2008	machine translation; statistical machine; named entity; semi supervised; large scale; conditional random; asr error; letter phoneme; phoneme conversion; language modeling; entity recognition; natural language; best hypotheses; image annotation; pattern clusters; data driven; random fields; sentence fusion; monolingual corpora; weakly supervised
2009	machine translation; statistical machine; semi supervised; coreference resolution; question answering; dependency parsing; part speech; grammar induction; named entity; phrase based; large scale; multi document; relation extraction; finite state; letter phoneme; phoneme conversion; speech recognition; spoken dialog; document summarization; hidden markov
2010	machine translation; statistical machine; semantic role; role labeling; context free; phrase based; cross lingual; dependency parsing; finite state; semi supervised; textual entailment; selectional preferences; fine grained; coreference resolution; correcting errors; spoken dialogue; free rewriting; part speech; word sense; rewriting systems
2011	machine translation; dependency parsing; statistical machine; semi supervised; context free; part speech; natural language; word alignment; markov models; speech tagging; dual decomposition; sentiment classification; cross lingual; search queries; hidden markov; sentiment analysis; large scale; parallel corpora; relation extraction; semantic role
2012	machine translation; statistical machine; cross lingual; semi supervised; named entity; large scale; dependency parsing; error correction; coreference resolution; phrase based; relation extraction; part speech; natural language; david chiang; head driven; hierarchical phrase; fine grained; finite state; language processing; label propagation
2013	machine translation; statistical machine; question answering; social media; semi supervised; cross lingual; part speech; word segmentation; natural language; distributional semantics; beam search; random walk; relation extraction; comparable corpora; coreference resolution; large scale; distant supervision; latent variable; pos tagging; word alignment
2014	machine translation; relation extraction; statistical machine; neural network; dependency parsing; cross lingual; recurrent neural; semi supervised; sentence level; word segmentation; distributional semantics; document summarization; distant supervision; domain adaptation; question answering; empirical study; similarity contextual; word sense; neural networks; weakly supervised

Table 3 – Top bigrams per year, extracted from publication titles and abstracts

[5] All stop words were removed prior to running frequency analysis on the content of the publications.

In each instance, the term *machine translation* is the most frequent topic by far and *statistical machine translation* is the most frequent trigram. To illustrate the importance of *machine translation*, consider the weighted scores listed in Table 4 where log likelihood was used to compute the top bigrams extracted from the 2013 titles only. As can be seen, machine translation plays a crucial role overall, since other top terms are related topics such as *phrase-based, statistical,* and *language modeling.* Another prominent topic across the years is parsing and in particular, *dependency parsing*.

machine translation:337.671716	mechanical turk:91.025780
statistical machine:190.548802	scale discriminative:87.637169
phrase based:119.247732	discriminative training:85.023513
boundary words:108.882638	translation tasks:84.597282
state art:96.746858	bleu points:83.258015
sparse features:93.487253	parallel data:78.643883

Table 4 - Top weigthed bigrams from 2013 titles

The next step was to identify themes that remained prominent throughout the years under study. For this purpose, we selected the terms from Table 3 that appeared among the top 20 ranked items for most of the years between 2007 and 2014 for comparison. The resulting top terms and their rank per year are illustrated in Figure 6. As can be seen, the term *machine translation* is the most frequently occurring bigram every year between 2007 and 2014, with *statistical machine* also remaining a constant frequently occurring term. Besides statistical machine translation, other topics occur in the top 20 terms across the years, but they do not have the same constancy. For instance, *named entity* peaks in 2008 but drops off in 2010-2011; *coreference resolution* seems to occur more often among the top ranks; and the use of *relation extraction* is on the rise since 2011.

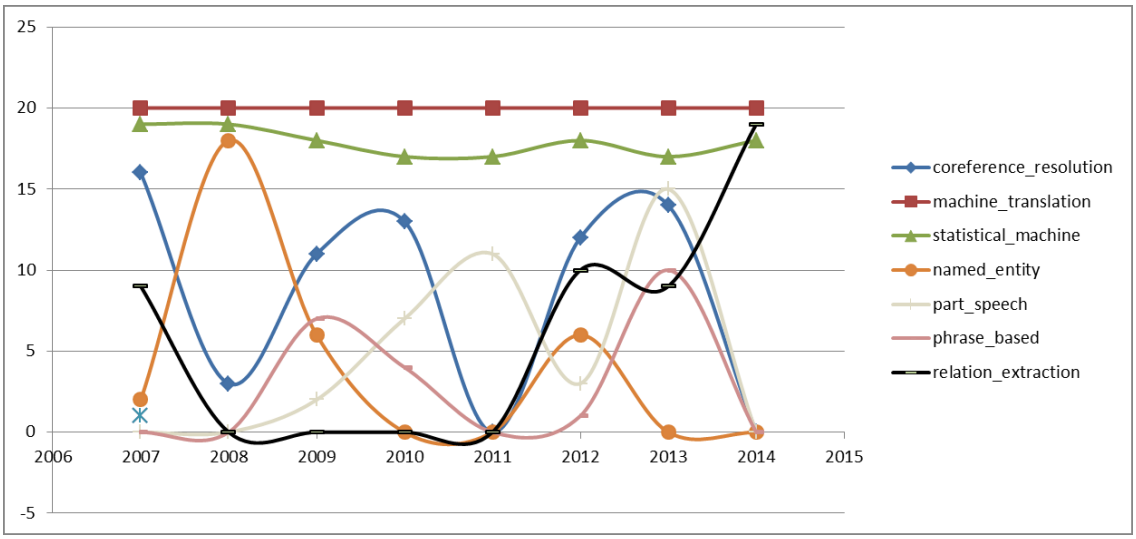

Figure 6 – Selected bigrams from top 20 frequent terms occurring 2007 through 2014 in the content of publications. The y-axis represents the rank of the term (20 being highest, 1 lowest, 0 means no occurrence that year).

Emerging topics

Although certain terms occur often between 2007 and 2014, we identified a number of terms as "new" terms appearing for the first time in the top 20 rank on a particular year. These terms, in Table 5, could

signify a new or emerging theme in the papers, which may or may not occur in future publications. For example, *social media* first appears in the top 20 terms in 2013 while *neural networks* first appears among the most frequent terms in 2014[6].

Year	Top "new" terms
2007	eye gaze, information retrieval, target language
2008	fine grained, large scale, pattern clusters, conditional random fields, penn treebank, bleu score, active learning, query expansion, log linear
2009	semantic role, hidden markov
2010	cross lingual, semantic role labeling, textual entailment
2011	dialogue act, sentiment classification, verbal feedback, word alignment
2012	error correction, gold standard, tense aspect, inter annotator
2013	social media, latent variable, pos tagging, distant supervision
2014	distributional semantics, neural network(s), document summarization, similarity contextual, weakly supervised, word sense, word segmentation

Table 5 - New terms per year from the top 20 frequent bigrams

The following figure looks at the occurrence of a few concepts that seem to be frequent in 2013 publications and compares their relative frequency in the last 7 years. As can be seen, a few concepts such as *crowdsourcing* and *social media* seem to be on the rise[7], while *machine translation* and *semantic* are frequent through the years.

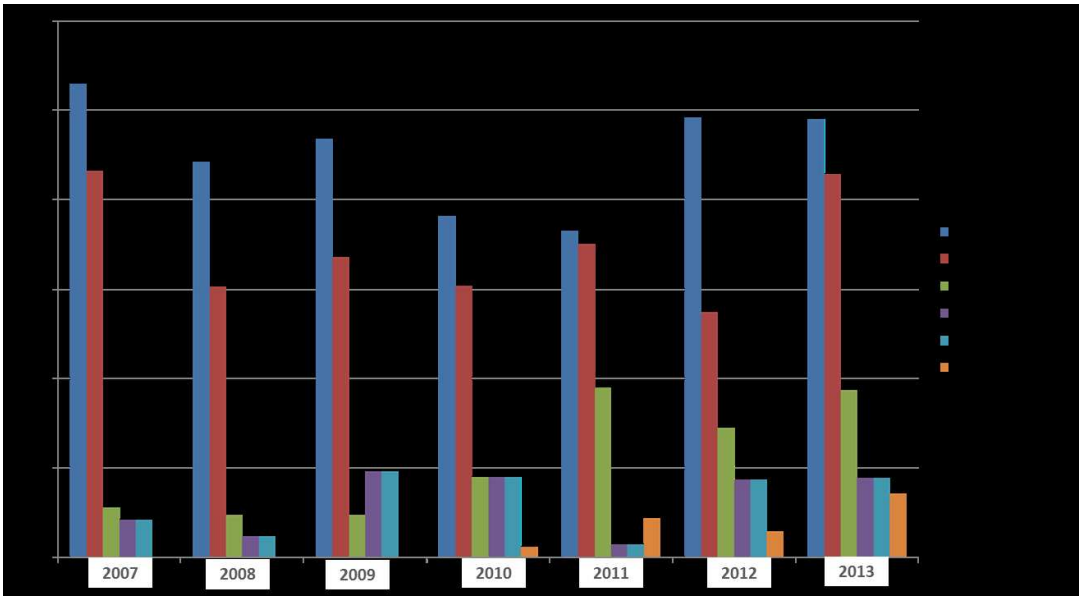

Figure 7 - Proportion of publications per year with mentions of concepts frequent in 2013

These results may suggest the rise of certain terms as "new", but in order to have a more systematic analysis of potentially emerging terms we ran GramReaper, a contrastive corpus linguistic tool that

[6] We selected the 2007 terms differently. We included them if they were not picked up in future publications.
[7] Figure 7 actually represents concepts rather than terms since terms like *Mechanical Turk* were counted as part of *crowdsourcing*, and *Twitter* was included within the *social media* category.

identifies statistically interesting differences between corpora[8]. Various statistical measures (chi-square, log-likelihood, tf-idf, and pointwise mutual information) were applied to detect terms that distinguish publications in 2013-2014 from all publications preceding 2013 (i.e., publications in 2007-2012). These analyses were run on different phrase sizes, ranging from unigrams to 4-grams, extracted from the titles of the publications. Table 6 shows the top terms that have been selected using the log-likelihood statistical measure as new terms appearing in 2013-2014 vs. terms appearing prior to 2013.

Term	Log-likelihood score
distributional semantics	18.9468942
social media	11.11950834
neural networks	7.868639036
extraction with	6.682700929
study on	6.136260816
relation extraction with	5.65185284
word segmentation	4.926923546
chinese word segmentation	4.692330884
relation extraction	4.309715184
distant supervision	4.073413458
beam search	3.477470562
language learning	3.477470562
random walk	3.477470562
chinese word	3.57215662
based translation model	2.93758749
case study	2.93758749
disfluency detection	2.93758749
domain independent	2.93758749
level discourse	2.93758749
for document	2.93758749
semantic models	2.93758749
shift reduce	2.93758749
tagging with	2.93758749
question answering	2.621232497
sentiment analysis	2.601670434

Table 6 - Top emerging terms in 2013 and 2014 publications, based on publication titles

Language Coverage

Figure 8 shows the coverage of distinct languages in the ACL and NAACL publications, based on the titles only. In addition to the languages shown, there are 36 papers covering language families based on typology (e.g., South African, Germanic, Indian, Turkic, Amerindian) and language characteristics (e.g., compounding, morphologically rich). There are also 3 paper titles focusing on *low-resource languages* and 13 papers discussing *dialects* or *dialect identification*.

[8] GramReaper was developed by Tim Allison at MITRE.

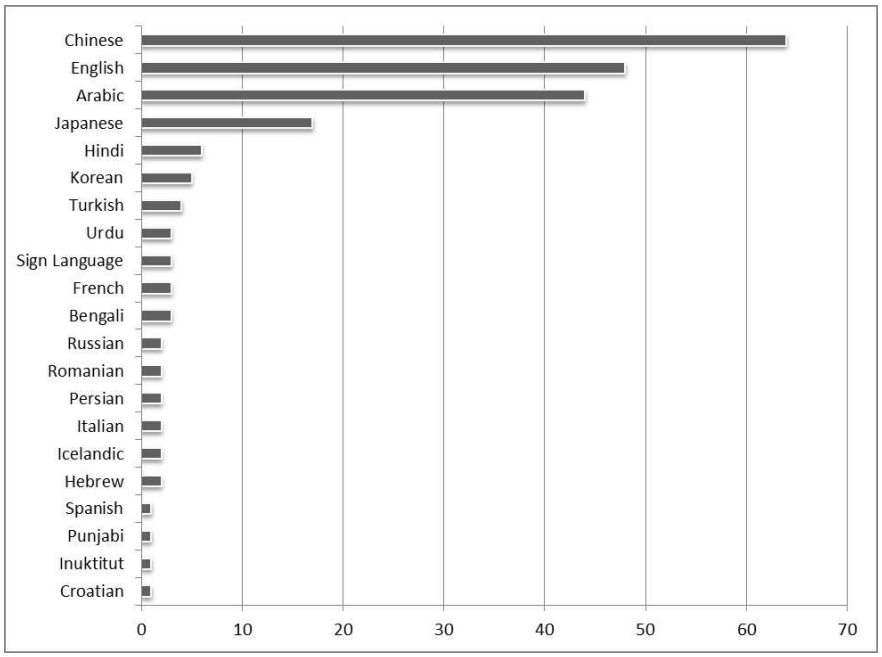

Figure 8 - Language coverage in ACL/NAACL titles, 2007-2014 (not an exhaustive list). *x*-axis represents document count – i.e., number of documents containing the language in the title.

6. Clustering

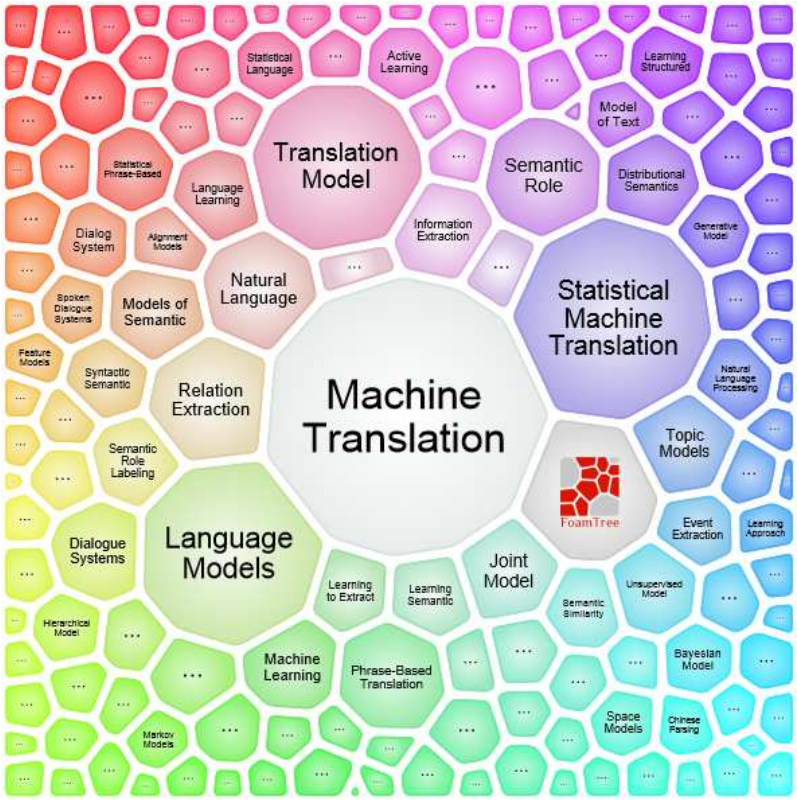

Figure 9 – Hierarchical structure of topic clusters using FoamTree

Another way of visualizing the data is to cluster the topics. We used the Carrot^2 open-source platform with the Lingo algorithm on the titles of publications to identify the main topic clusters. The key characteristic of the Lingo algorithm offered by Carrot^2 is that it first identifies the labels for the clusters (by building a term document matrix) and then assigns documents to the labels (if the documents contain the label's words) to form the final clusters. The results clearly confirm the central role of *machine translation* and related topics in the ACL and NAACL publications as illustrated in Figure 9.

Figure 10 shows the relationships between the various topic clusters. Again, this visualization clearly shows how machine translation is fueling much of the research in the field.

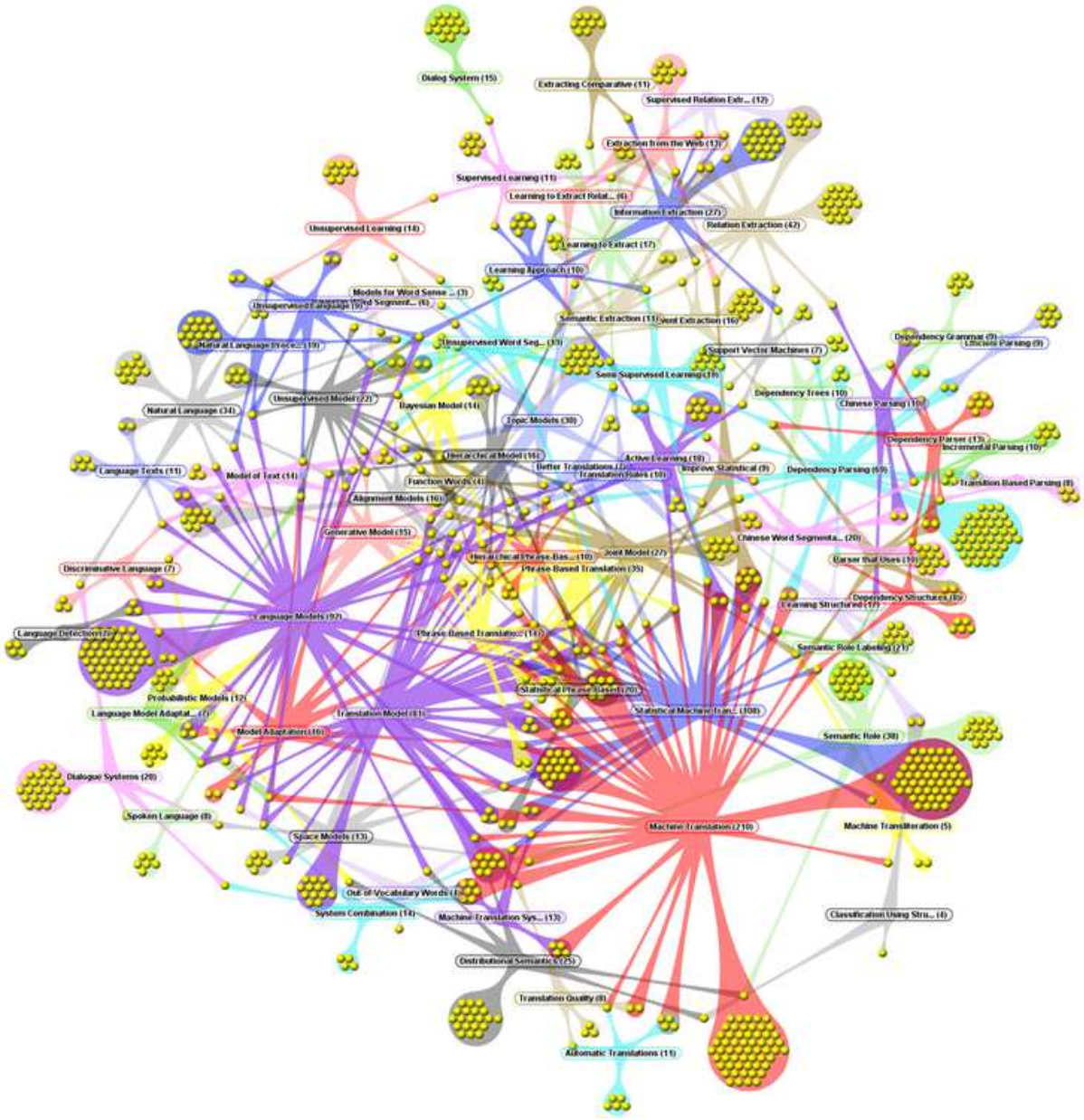

Figure 10 - Interrelations between topic clusters using Aduna visualization

Figure 11 highlights the relations between Machine Translation and other topics – the graph shows that MT is closely related to the following topics: translation model (46 out of 81 documents), model adaptation (4/16), language models (10/92), statistical phrase-based (17/20), phrase-based translation (20/49), active learning (3/18), dependency structures (1/8), semantic role labeling (1/21), etc. It is however not closely related to the various clusters seen in the upper part of Figure 10, which are labeled as information extraction (relations, event, semantic) and word sense disambiguation.

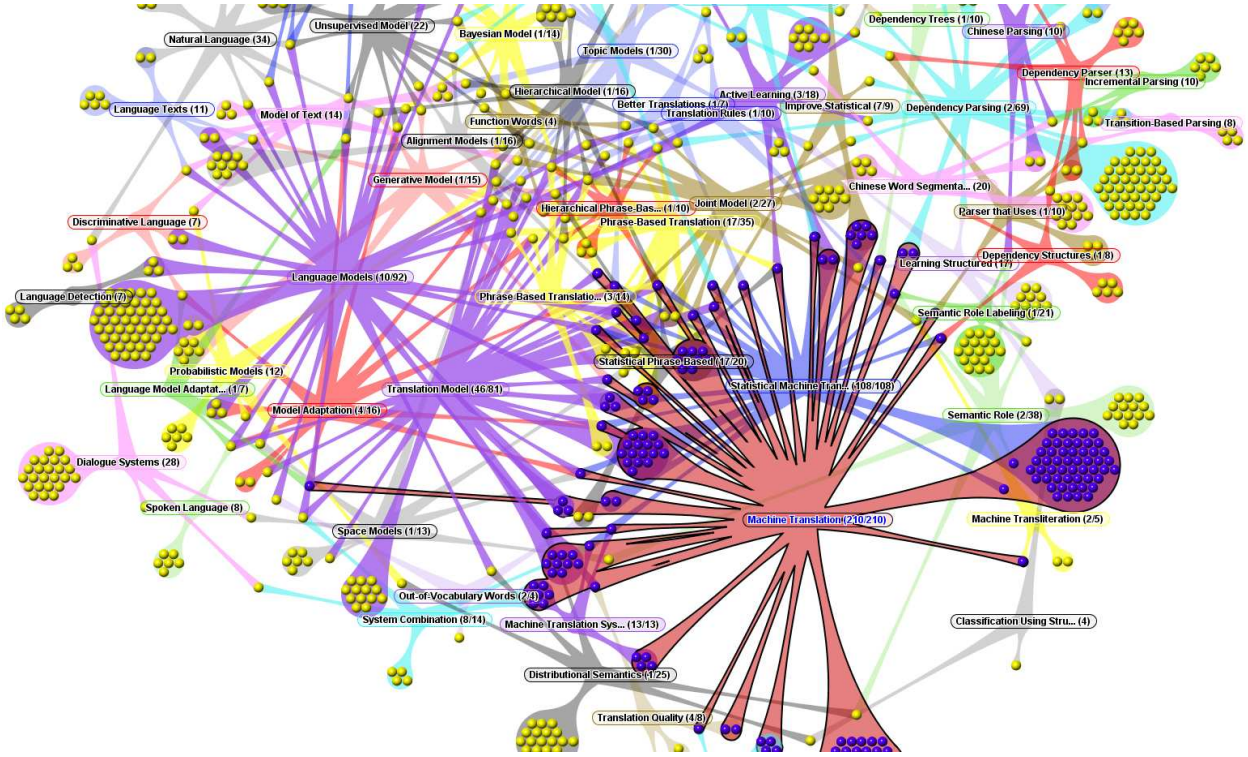

Figure 11 - Interrelations between the Machine Translation cluster and other topics, using Aduna visualization

7. Co-Authorship Network

Network Overview

We extracted ACL and NAACL publication authors and we used them to build a collaboration network, where the nodes represent authors and the edges indicate a co-author relationship. Thus, every time a person co-authors a paper with another author, we recorded an edge between the two authors. In the case of multi-authored papers, we created edges between all two members of the authors list. We investigated the properties of the co-authorship network by (a) contrasting the network characteristics and central nodes for each year to explore changes in the network through time, (b) studying the characteristics of the full network, combining the nodes and edges for 2007-2014, and (c) exploring the giant connected component in the network for 2007-2014 to identify the central authors. We used the Gephi[9] network analysis and visualization software in all analyses.

[9] http://gephi.github.io/

The collaboration network is a weighted undirected graph. The network is undirected since co-authorship is a two-way relationship, and the weights associated with the edges represent the number of times two persons have co-authored in the dataset. The full network, combining all nodes and edges from 2007-2014, consists of 3184 nodes (authors) and 8171 edges (co-author relations).

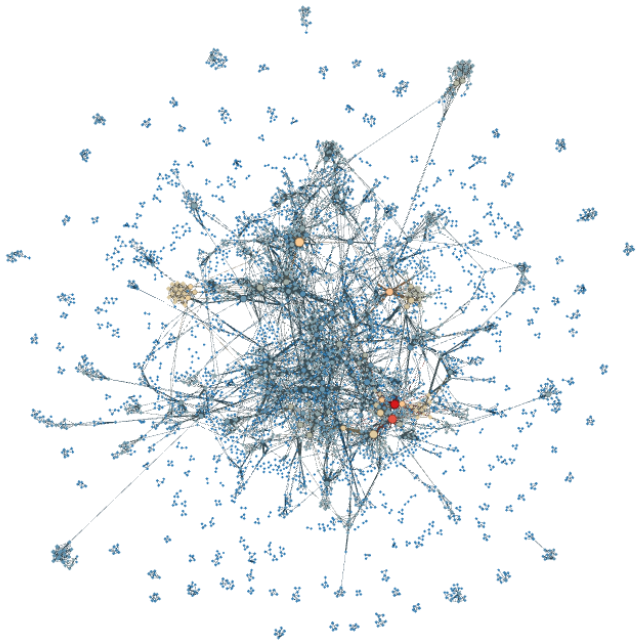

Figure 12 - Full collaboration network of ACL and NAACL publications, 2007-2014. Red nodes indicate higher degree.

It should be noted that we did not edit or resolve the name variants in the ACL and NAACL publications for the purpose of this study. Thus, the different variants of an author's name – e.g., with or without middle initial, misspellings or variants on diacritics, different transliterations – are considered as distinct nodes which will result in some level of error.

Slight rise in collaborations

Table 7 shows the number of nodes and edges within the network for each year.

	2007	2008	2009	2010	2011	2012	2013	2014
Nodes	763	469	936	912	753	789	1275	432
Edges	1407	761	1796	1655	1294	1499	2583	874

Table 7 - Distribution of ACL and NAACL nodes (authors) and edges (collaborations) per year, 2007-2014.

The figures below show the total count of nodes and edges per year (left figure) and the proportion of nodes and edges with respect to the number of publications (right figure). The latter shows that the relative proportion of nodes (authors) in the network remains constant throughout the years, although there seems to be a slight increase in the number of edges (collaborations).

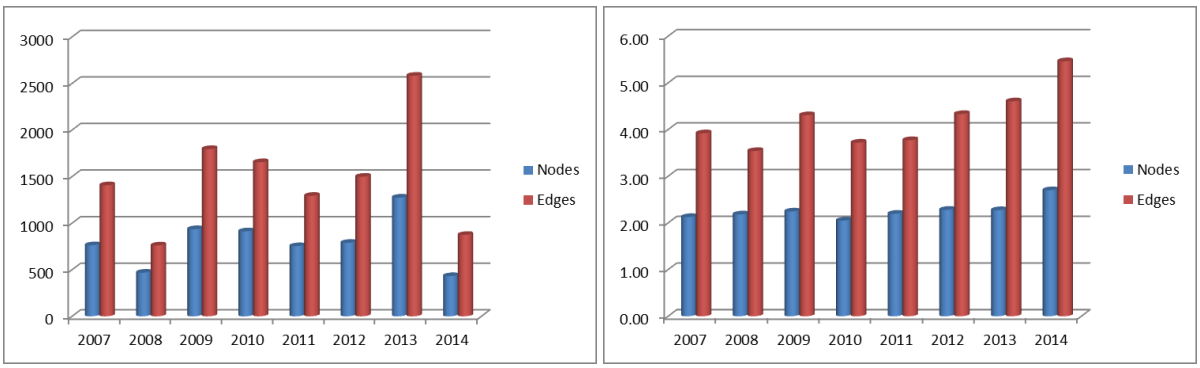

Figure 13 - Count (left) and relative proportion to number of publications (right) of nodes and edges per year.

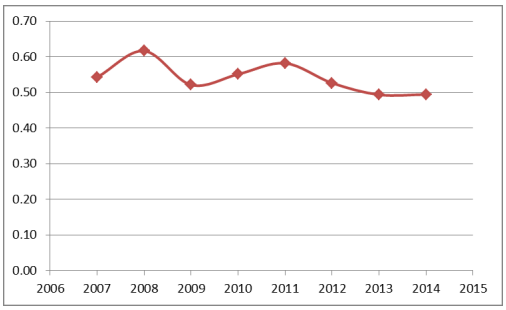

Figure 14 - Relative number of nodes (authors) to edges (collaborations) per year

Similarly, Figure 14 illustrates the relative number of nodes to edges per year and shows that the ratio remains mostly constant, although there is a slight decrease in recent years indicating that there may be more collaborations taking place in the field[10]. It is interesting to note that the 2008 ACL data included a higher number of isolated nodes, indicating more single-authored papers.

These analyses show that the ACL/NAACL collaboration network is a very cohesive network with a large central network dominating the publication scene –this giant component includes 2658 or 83% of the total 3184 nodes of the full network. The second largest component is far smaller (about 1% of full network). This densely interconnected giant component behavior is a characteristic signature of networks that are well inside what Newman refers to as the *transition percolation* regime, which is itself a property of many scientific collaboration networks.

> *In networks with very small numbers of connections between individuals, all individuals belong only to small islands of collaboration or communication. As the total number of connections increases, however, there comes a point at which a giant component forms – a large group of individuals who are all connected to one another by paths of intermediate acquaintances. […] almost everyone in the community is connected to almost everyone else by some path (probably many paths) of intermediate coauthors. […] It appears that scientific collaboration networks are not on the borderline of connectedness—they are very highly connected and in no immediate*

[10] This result is not very significant but it would be in line with previous research claiming that academic collaboration is on the rise and single-authored papers are becoming less common across the sciences (Newman 2001, Rawlings and McFarland 2011).

danger of fragmentation. (Newman 2001).

	2007	2008	2009	2010	2011	2012	2013	2014	Full Network	Giant Comp.
Nodes	763	469	936	912	753	789	1275	432	3184	2658
Edges	1407	761	1796	1655	1294	1499	2583	874	8171	7181
Avg Degree	3.688	3.245	3.838	3.629	3.437	3.8	4.052	4.046	5.133	5.403
Avg Weighted Degree	4.524	4.772	4.998	5.037	4.922	5.323	5.22	5.602	5.932	6.334
Network Diameter	17	11	16	14	15	14	13	14	16	16
Graph density	0.005	0.007	0.004	0.004	0.005	0.005	0.003	0.009	0.002	0.002
Connected components	150	86	142	168	118	112	196	54	183	1
Avg. clustering coefficient	0.748	0.675	0.715	0.718	0.688	0.734	0.765	0.763	0.827	0.874
Avg. path length	6.11	5.354	5.998	5.687	5.991	6.114	6.025	6.053	6.253	6.373

Table 8 - Network analysis results for all collaboration networks

Table 8 provides the network analysis results:

- ***Average degree:*** Average number of co-authorship relations per node.
- ***Average weighted degree:*** Average number of co-authorship relations per node, taking into account recurrent co-author relations (represented as weight on the edges).
- ***Network diameter:*** Diameter of a graph is defined as the length of the longest shortest path between any two nodes. It indicates distance between the two most distant nodes.
- ***Graph density:*** The density of the graph indicates how close the network is to complete. A complete graph has all possible edges (i.e., a fully connected network) and density is equal to 1.
- ***Connected components:*** The total number of connected components indicates the various clusters found in the larger network. For instance, the giant component consists of only a single large connected component, while in the full network there are also 182 smaller clusters of co-authorship or smaller islands of collaboration (most representing a single publication).
- ***Average clustering coefficient:*** Clustering coefficients are used to determine whether a network can be labeled as a small-world network. A real-world network will generally have a much higher clustering coefficient than a random network of the same size.
- ***Average path length:*** In a network, the average smallest number of steps along edges between any two nodes is called the *average shortest path*.

As already noted, the giant connected component includes about 83% of the authors in the dataset, with about 17% falling outside of the giant component. The average path length between all pairs of authors in our dataset for whom a connection exists is about six in all networks[11]. In other words, there are 6 degrees of separation in the field of ACL/NAACL computational linguistics – this also reflects the

[11] The average path length and network diameter are measured based on the giant component in the network.

findings for scientific publications in general (Newman 2001). The clustering coefficient probes for the existence of clusters (local communities in which a higher than average number of people know one another). This value equals 1 for a fully connected graph. The average clustering coefficient for the networks in Table 8 shows that there is a very strong clustering effect, which has been argued to signal that new collaborations are regularly brokered in the field. These results show that the networks exhibit *small-world network* characteristics (Watts and Strogatz 1998).

This conclusion is confirmed by the power law characteristic of degree distribution in the network, indicating a preference for edge attachment to a small number of high degree nodes[12]. All of these results parallel the previous research on the field of Computational Linguistics (e.g., Radev et al 2009).

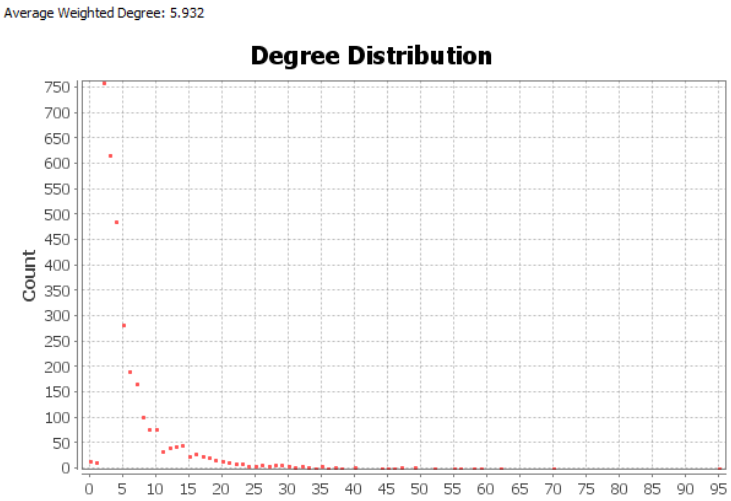

Figure 15 - Degree distribution in the full network; *x*-axis is degree value and *y*-axis is the number of nodes (authors) with that degree. The distribution shows a few authors with very high collaboration degree and many authors with very small co-authorships. (*source:* Gephi)

The existence of a giant component with a small average path length has been argued to allow news of important discoveries and scientific information to reach most members of the network and information typically circulates much faster in such networks. The small world network properties also signal a rather robust network that cannot easily be fragmented. At the same time, small world networks have been found to be more resistant to change and innovation due to the tendency for preferential attachment (Steen et al 2010). However, it should be recognized that about 17% of the authors in the field are disconnected from the network with potential consequences that could further be studied.

The Giant Component: The Central Role of Machine Translation

In what follows, we discuss analyses performed on the giant component of the full network. The central role that machine translation, and in particular statistical machine translation occupies in the ACL and NAACL publications can also be seen in the collaboration network. Figure 16 shows the main communities detected using the Louvain Modularity algorithm on the 2007-2013 data; clusters were

[12] For this study, tests were not run to ascertain that the ACL/NAACL networks exhibit power law characteristics, as it is beyond the scope of this paper, but see Radev et al (2009) for a discussion.

then manually tagged for main topic based on the publication titles. The largest cluster overall and the second important cluster are both focused on *statistical MT*. The smaller and related clusters also involve statistical MT in some way, as in developing *MT Resources* or *MT Tools*, focusing on specific theoretical elements related to statistical MT such as improvements in *learning theory*, and applications of statistical MT in other languages such as Japanese and Chinese. Within the periphery, other topics or subject areas are investigated such as *lexical semantics, extraction,* and *discourse.*

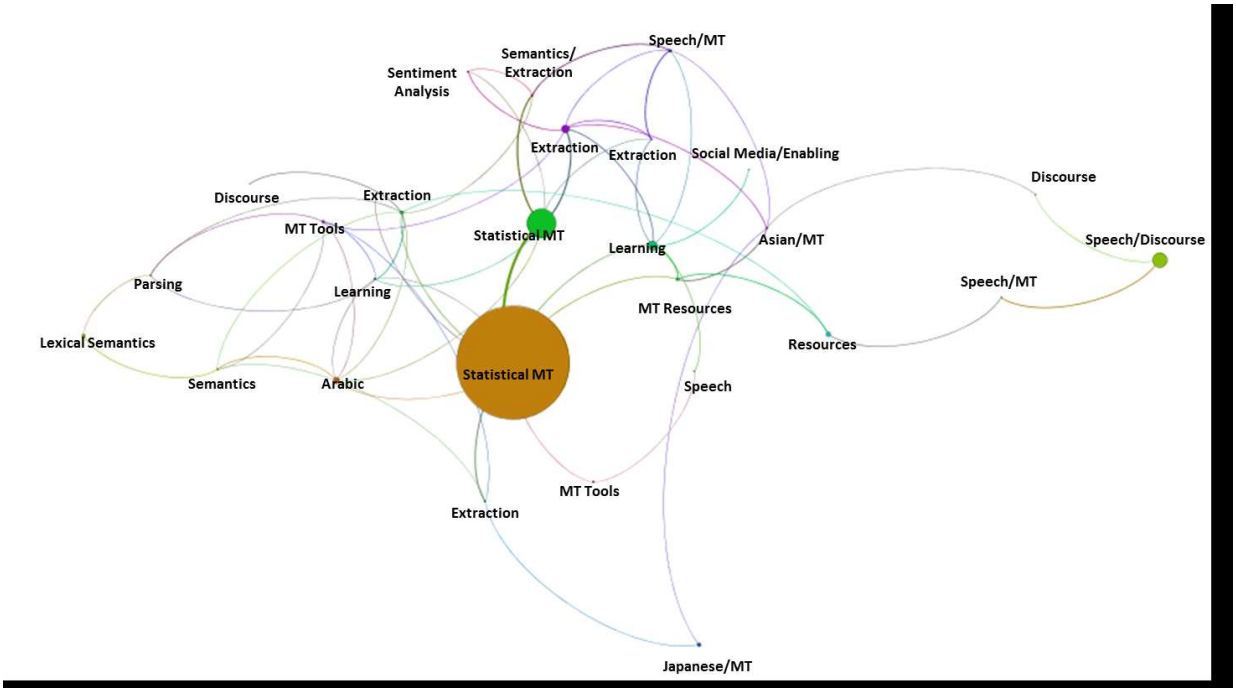

Figure 16 - Main clusters in the ACL and NAACL co-authorship network (2007-2013)

Figure 17 shows the various communities computed using the modularity algorithm for all publications from 2007 through 2014. The major communities were again manually tagged for topic or subject area based on the titles of publications. These clusters identify distinct communities that closely co-author based on topic, but also based on institution of affiliation and associated country, or because of a common language of analysis.

As already discussed, statistical MT is the most dominant subject area category and fuels other areas of research such as resource development and advances in parsing and learning algorithms. Statistical MT covers the largest and most central clusters in the graph. In addition, statistical MT is the main focus of research in the communities at the upper part of the graph where the emphasis is on Chinese and Japanese languages. The other important area is knowledge discovery and extraction, and different interrelated communities exist with different approaches to the topic. The communities working on lexical semantics, discourse or ASR seem to be at the periphery of the network.

This cluster graph displays a rather interconnected and interdisciplinary (within existing areas) field of research, with the caveat that most research is fueled by applications for machine translation.

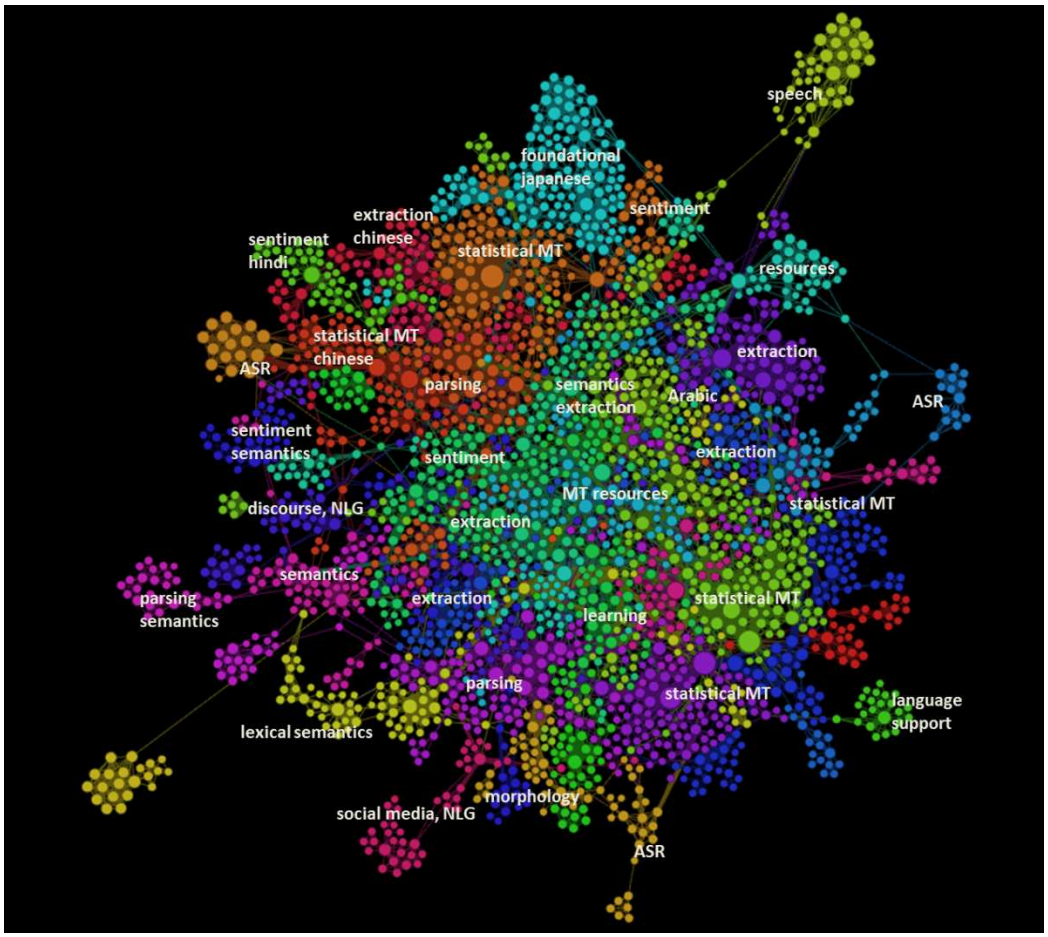

Figure 17 - Clusters in the collaboration network, giant component, 2007-2014

Central Authors in the Giant Component

Node metrics are provided in Table 9 with the top 20 ranked authors in several centrality categories. The nodes with the highest *Degree centrality* are authors who collaborate most often with others. *Weighted Degree* takes into account recurrent co-authorships. In other words, authors with high weighted degree that do not appear in the top 20 with highest degree (e.g., Mu Li) are those who tend to collaborate with the same authors. *Betweenness centrality* measures the extent to which a particular point lies on shortest paths between the various other points in the graph. These authors may play an important intermediary role within the network, serving as a bridge connecting otherwise disparate authorship clusters. For example, Pushpak Battacharyya's cluster is connected to the giant component only through Trevor Cohn (via Gholamreza Haffari). The intermediary role of Trevor Cohn is also reflected in the breadth of topics apparent in his publication titles, shown in Table 10. In addition, the top 20 authors with high *Eigenvector centrality* are those considered important and central by virtue of being

connected to other nodes that are central within the network. *Closeness centrality* measures the shortest path between a node and all the other nodes in the network. This measure allows us to find the globally central node in the network, as the more central a node is the lower its total distance to all other nodes. These authors hold an important role in spreading information to all authors. Table 9 provides the list of the top 20 authors with the smallest closeness values.

	Weighted Degree	Degree	Betweenness	Eigenvector	Closeness
1	Ming Zhou	Ming Zhou	Chris Dyer	Chris Callison-Burch	Chris Dyer
2	Noah Smith	Chris Dyer	Trevor Cohn	Chris Dyer	David Chiang
3	Chris Dyer	Chris Callison-Burch	Heng Ji	Heng Ji	Vladimir Eidelman
4	Chris Callison-Burch	Noah Smith	Kevin Knight	Philipp Koehn	Kevin Knight
5	Qun Liu	Heng Ji	Alessandro Moschitti	Jonathan Weese	Phil Blunsom
6	Dan Klein	Qun Liu	Giorgio Satta	Ming Zhou	Trevor Cohn
7	Regina Barzlay	Mark Dredze	Vladimir Eidelman	Yao Meng	Yoav Goldberg
8	Heng Ji	Yang Liu	Ming Zhou	Fuliang Weng	Philip Resnik
9	Yang Liu	Regina Barzlay	Yang Liu	Mark Dredze	Philipp Koehn
10	Mu Li	Kevin Knight	Mark Dredze	Dipanjan Das	Chris Quirk
11	Min Zhang	Pushpak Battacharyya	Chin-Yew Lin	Noah Smith	Dan Klein
12	Ting Liu	Ralph Grishman	Yoav Goldberg	Wade Shen	John DeNero
13	Ido Dagan	Ting Liu	David Chiang	Hieu Hoang	Hendra Setiawan
14	Haizhou Li	Eduard Hovy	Mona Diab	Matthew Purver	Heng Ji
15	Mark Dredze	Tiejun Zhao	Timothy Baldwin	Sebastian Varges	Joakim Nivre
16	Dan Roth	Haizhou Li	Martha Palmer	Richard Zens	Adam Lopez
17	Kevin Knight	Philipp Koehn	Chris Quirk	Alexandra Birch	Liang Huang
18	Eduard Hovy	Mark Johnson	Eduard Hovy	Marcello Federico	Dipanjan Das
19	Ralph Grishman	Joakim Nivre	Philipp Koehn	Ralph Grishman	Karl Moritz Hermann
20	Chris Manning	Stephan Vogel	Wen Wang	Stanley Peters	Jonathan Weese

Table 9 - Node metrics for collaboration network's giant component

Reducing Annotation Effort for Quality Estimation via Active Learning
QuEst - A translation quality estimation framework
Inducing Compact but Accurate Tree-Substitution Grammars
Inducing Synchronous Grammars with Slice Sampling
Evaluating a Morphological Analyser of Inuktitut
Machine Translation by Triangulation
A Discriminative Latent Variable Model for Statistical Machine Translation
A Gibbs Sampler for Phrasal Synchronous Grammar Induction
A Note on the Implementation of Hierarchical Dirichlet Processes
Blocked Inference in Bayesian Tree Substitution Grammars
A Hierarchical Pitman-Yor Process HMM for Unsupervised Part of Speech Induction
Modelling Annotator Bias with Multi-task Gaussian Processes
A Markov Model of Machine Translation using Non-parametric Bayesian Inference
An Infinite Hierarchical Bayesian Model of Phrasal Translation
A user-centric model of voting intention from Social Media

Table 10 - Publications by Trevor Cohn, an author with high betweenness centrality value

Appendix 4 provides the top nodes per year. It is interesting to note that, although there are some

fluctuations in the first years of the dataset, the top authors in the collaboration network remain rather constant from year to year in more recent years (i.e., 2010-2014), with the same main nodes playing a central role in terms of degree and betweenness.

Central Institutions in the Giant Component

The institutions of the authors central to the network are mainly focused in the U.S. and China (Table 11), while some of the authors that are central as intermediary nodes within the network are affiliated with institutions in Europe (Table 12).

Author	Institution	Country
Ming Zhou	Microsoft Research Asia	China
Chris Dyer	Carnegie Mellon University	USA
Chris Callison-Burch	Johns Hopkins University	USA
Noah Smith	Carnegie Mellon University	USA
Heng Ji	Rensselaer Polytechnic Insitute (RPI)	USA
Qun Liu	Chinese Academy of Sciences	China
Mark Dredze	Johns Hopkins University	USA
Yang Liu	Tsinghua University	China
Regina Barzlay	MIT	USA
Kevin Knight	Information Sciences Institute (ISI), USC	USA
Pushpak Battacharyya	Indian Institute of Technology	India
Ralph Grishman	NYU	USA
Ting Liu	Harbin Institute of Technology	China
Eduard Hovy	Carnegie Mellon University	USA
Tiejun Zhao	Harbin Institute of Technology	China
Haizhou Li	Nanyang Technological University	Singapore
Philipp Koehn	University of Edinburgh, Johns Hopkins	UK, USA
Mark Johnson	Macquarie University	Australia
Joakim Nivre	Uppsala University	Sweden
Stephan Vogel	Carnegie Mellon University	USA

Table 11 - Institutions of central authors with high degree values

Author	Institution	Country
Trevor Cohn	University of Sheffield (senior lecturer)	UK
Alessandro Moschitti	University of Trento	Italy
Giorgio Satta	University of Padua	Italy
Vladimir Eidelman	University of Maryland (PhD candidate)	USA
Chin-Yew Lin	Microsoft Research Asia	China
Yoav Goldberg	Ben-Gurion University (senior lecturer)	Israel
David Chiang	Information Sciences Institute (ISI), USC	USA
Mona Diab	George Washington University	USA
Timothy Baldwin	University of Melbourne	Australia
Martha Palmer	University of Colorado Boulder	USA
Chris Quirk	Microsoft Research	USA
Wen Wang	SRI	USA

Table 12 - Institutions of central authors with high betweenness values

Also see Appendix 5 for a case study on the 2013 network components focused on machine translation

and parsing.

8. Discussion

This report provides topic and collaboration network analyses performed on ACL and NAACL publications from 2007 through 2014. We were able to detect important domains and subject areas within the field, identify the central authors and institutions, and present results on changes in topic or term trends in the recent years within the field. There have been several studies based on the ACL datasets; some of these results are discussed in this section.

The collaboration network analysis results we obtained parallel those found in Radev et al (2009a, 2009b) which they obtained by studying the full ACL anthology. These authors also show that the collaboration and citation networks can be identified as small-world networks indicating that the networks are very well connected and that there is active collaboration in the field. In addition, there are a small number of papers or authors which are attracting the majority of collaborations; these authors play a very strong role in the overall structure of the network. Radev et al did not consider the change in the network, however. For a large-scale network analysis of ACL authors and publications, the reader is referred to the *ACL Anthology Network* at http://clair.eecs.umich.edu/aan/.

Paul and Girju (2009) perform a contrastive topic model study of a subset of the computational linguistics and formal linguistics publications. They detect *text classification* and *statistical/probabilistic methods* have increased as topics of study in HLT, while *formal semantics, natural language interfaces* and *speech act interpretation* have declined. However, several subject areas such as *word sense disambiguation, semantic role labeling,* and *event/temporal semantics* are on the rise. Moreover *morphology, prosody* and *quantifiers* have seen a steady decline. These conclusions are in line with the findings in this paper. It is however interesting that the authors do not mention a rise in machine translation specifically, given how dominant this subject area has become in the current field.

Sim et al (2012) perform graph clustering on collaboration and citation networks of authors to identify the topic factions in the field, based on the 500 top cited authors. The results obtained differ quite a bit from the findings in this current study. In particular, the largest faction is the word sense disambiguation (WSD) faction (42 authors), followed by formalisms (31) and discourse (29). There are two machine translation factions, each containing 9 authors. The differences obtained could be due to the way the papers were categorized – in our current study, any paper that specifically mentioned MT was tagged in the machine translation category, including papers dealing with MT evaluation and WSD specifically designed for machine translation. In the Sim et al study, MT and MT evaluation are kept in distinct categories. In addition, the authors treat parsing as its own distinct faction, separate from the formalisms faction which includes parsing-related work.

The work in Johri et al (2011) is very different from the current study, however, since these authors investigate the types of collaborations embodied by differing levels of seniority and contribution from each co-author. These authors also study “author signatures” – identifying main terms by author using Labeled LDA algorithm. Although the latter was not the focus of this current study, the results suggest

differences in authors based on their role in the network (central hub vs. bridging node) which coincides with how diversified they are in their signature.

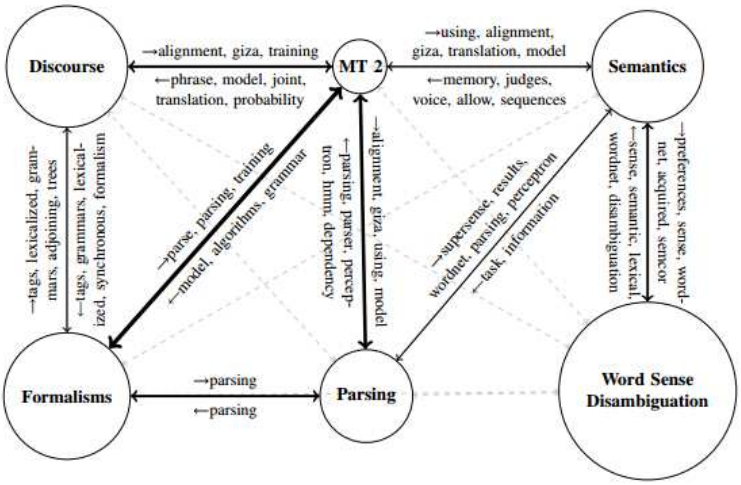

Figure 18 - Citations among some factions. Node size reflects faction size; edge thickness reflects number of inter-faction citations. Words on the edges are the highest weighted words. (*source:* Sim et al 2012)

Several improvements can be made to this report by cleaning up some of the data and algorithms to remove any sources of error. For instance, the author names with name variants can be resolved to capture a more accurate social network or the classification model can be tuned to obtain better results. The analysis can also be extended by including the full content of the publications. Nevertheless, the overall trends observed through network analysis as well as through application domain and subject area detection provide interesting conclusions about the field and how it is changing.

References

N. Johri, D. Ramage, D. A. McFarland, D. Jurafsky (2011). A study of academic collaboration in computational linguistics with latent mixtures of authors. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.*

M. E. J. Newman (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Science,* 98:404-409.

M. Paul and R. Girju (2009). Topic modeling of research fields: an interdisciplinary perspective. *Proceedings of RANLP'09.*

D.R. Radev, M. Joseph, B. Gibson and P. Muthukrishnan (2009a). A bibliometric and network anlaysis of the field of computational linguistics. *Proceedings of JASIST.*

D.R. Radev, P. Muthukrishnan and V. Qazvinian (2009b). The ACL anthology netowrk corpus. *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries,* NLPIR4DL'09.

C. M. Rawlings and D. A. McFarland (2011). Influence flows in the academy: Using affiliation networks to assess peer effects among researchers. *Social Science Research,* 40(3):1001-1017.

Y. Sim, N. A. Smith, and D. A. Smith (2012). Discovering factions in the computational linguistics community. *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries.*

J.Steen, S. MacAulay and T. Kastelle (2010). A review and critique of the small world hypothesis: The best network structure for innovation? *Proceedings of Druid Summer Conference on Opening Up Innovation: Strategy, Organization and Technology.*

D. J. Watts and S. H. Strogatz (1998). Collective dynamics of 'small-world' networks. *Nature* 393:440-442.

Appendix 1: HLT Classification

Foundational Technology		
Name	**Definition**	**Tag**
Annotation / Data Management	Data is essential for language technology, and tools are required to collect, clean, annotate, and format data resources. *Includes lexicon development.*	AnDM
Enabling Component Technologies	Language processing capabilities such as word segmentation and part of speech tagging are used to build or enhance language technologies. *Includes morphology, POS taggers, parsers, word sense disambiguators, etc.*	ECT
Machine Translation	Machine translation technology translates content from one language into another	MT
Text-to-Speech / Speech Generation	TTS systems convert text into speech sounds.	TTS
Discourse and Rhetoric	*Analysis of discourse features, including coreference resolution, conversation analysis, dialogue systems, etc.*	DR
Theory	*Exploration of new theories and approaches (e.g., Deep Learning, Learning algorithms), without a specific application*	Th

Knowledge Discovery		
Name	**Definition**	**Tag**
Concept Extraction	Tools can determine which words represent objects or ideas, and those concepts can be described, related, and associated with attributes in an ontology. *Includes research on inference modeling and deception detection.*	CE
Entity Extraction	The names of people, places, organizations and other significant expressions such as dates and times can be recognized in speech or text.	EE
Relation and Event Extraction	Relations between entities such as members of organizations or participants in events can be identified in speech or text. Includes identification of temporal relations.	RE
Information Retrieval	Capabilities to search for language content of interest to the user range from keyword search to question answering and include cross-language information retrieval (CLIR)	IR
Name-Matching and Identity Resolution	Systems can recognize variations in names and other biographic information to improve search and screening	NM
Sentiment Analysis	Sentiment analysis classifies text as positive, negative or neutral and is usually used to extract and summarize opinions	SA
Summarization	Tools can produce brief representations or summaries of text content.	Sum
Knowledge Generation	*Generation of expressions, concepts, sentences for Natural Language Understanding or captioning applications.*	NLG

Language Professional Support		
Name	**Definition**	**Tag**

Error Correction	*Tools used to identify errors and provide automated correction to text*	EC
Cognitive	*Investigations on processing and efficiency analysis*	Cg
Translation Memory and Computer Aided Translation	Translation memory systems use previously translated sequences of language to increase translator efficiency and consistency. Computer assisted translation systems (CAT) use translation memory and other tools, such as spelling and grammar checkers or terminology databases, to facilitate the translation process.	TM

Triage		
Name	**Definition**	**Tag**
Optical Character Recognition / Image-to-Text	OCR systems convert images of written language to text – *especially in triage applications*	OCR
Transliteration / Phonetic Transcription	Tools are available to convert text from one writing system to another or from speech to representations of the speech sounds	Tr
Speaker and Author Attribute Detection	Technology is emerging to identify attributes of authors and speakers such as age, gender, and dialect.	AA
Language Identification	Tools are able to recognize the language that is used in speech or text data.	LID
Speaker and Author Identification	Systems can identify speakers and authors based on audio signals, handwriting, and/or linguistic style.	AID
Topic Detection	Topics can be identified within and across documents. *Includes classification*.	TD

Appendix 2: Top 30 words associated with application domains

Probability of the term given the class in the classifier model used to categorize the application domains.

	Foundational Technology	Knowledge Discovery	Triage	Language Professional Support
1	translation	extraction	topic	language
2	machine	summarization	classification	acquisition
3	machine translation	semantic	text	learning
4	parsing	learning	data	analysis
5	learning	relation	learning	method
6	word	sentiment	semantic	method for
7	language	entity	model	analysis of
8	dependency	relation extraction	models	citation
9	model	event	based	systems
10	models	inference	latent	correction
11	statistical	based	unsupervised	correction using
12	semantic	analysis	clustering	error
13	statistical machine	question	languages	error correction
14	statistical machine translation	opinion	documents	error correction using
15	for statistical	joint	topic models	language acquisition
16	for statistical machine	automatic	language	reading
17	segmentation	relations	identification	semantic
18	data	model	user	model
19	features	models	supervised	models
20	text	distributional	bayesian	approach
21	word segmentation	answering	topics	generation
22	unsupervised	extraction with	text classification	word
23	training	language	correction	text
24	lexical	data	extraction	context
25	improving	semantics	words	approach to
26	modeling	words	twitter	probabilistic
27	tagging	information	features	robust
28	neural	approach	based on	predicting
29	dependency parsing	online	extraction from	discriminative
30	alignment	generation	social	evaluation

Appendix 3: Top 20 bigrams ranked per year

2007		2008		2009	
Rank	Term	Rank	Term	Rank	Term
1	machine_translation	1	machine_translation	1	machine_translation
2	statistical_machine	2	statistical_machine	2	experimental_results
3	state_art	3	named_entity	3	statistical_machine
4	natural_language	4	semi_supervised	4	state_art
5	coreference_resolution	5	experimental_results	5	semi_supervised
6	paper_presents	6	state_art	6	paper_presents
7	experimental_results	7	fine_grained	7	semantic_role
8	eye_gaze	8	large_scale	8	question_answering
9	speech_recognition	9	pattern_clusters	9	sense_disambiguation
10	question_answering	10	conditional_random	10	coreference_resolution
11	sense_disambiguation	11	random_fields	11	hidden_markov
12	relation_extraction	12	paper_presents	12	training_data
13	semi_supervised	13	training_data	13	speech_recognition
14	non_projective	14	penn_treebank	14	phrase_based
15	domain_adaptation	15	bleu_score	15	named_entity
16	information_retrieval	16	active_learning	16	context_free
17	target_language	17	query_expansion	17	large_scale
18	paper_describes	18	coreference_resolution	18	results_show
19	named_entity	19	log_linear	19	part_speech
20	language_processing	20	results_show	20	natural_language

2010		2011		2012	
Rnk	Term	Rnk	Term	Rank	Term
1	machine_translation	1	machine_translation	1	machine_translation
2	state_art	2	state_art	2	state_art
3	experimental_results	3	natural_language	3	statistical_machine
4	statistical_machine	4	statistical_machine	4	natural_language
5	cross_lingual	5	dependency_parsing	5	cross_lingual
6	natural_language	6	semi_supervised	6	error_correction
7	role_labeling	7	paper_presents	7	large_scale
8	coreference_resolution	8	large_scale	8	paper_presents
9	semi_supervised	9	experimental_results	9	coreference_resolution
10	textual_entailment	10	part_speech	10	semi_supervised
11	context_free	11	dialogue act	11	relation_extraction
12	semantic_role	12	context_free	12	gold_standard
13	fine_grained	13	language_processing	13	fine_grained
14	part_speech	14	experiments_show	14	language_processing
15	dependency_parsing	15	cross_lingual	15	named_entity
16	paper_presents	16	sentiment_classification	16	tense_aspect

17	phrase_based	17	training_data	17	inter_annotator
18	question_answering	18	verbal_feedback	18	part_speech
19	results_show	19	word_alignment	19	experimental_results
20	data_sets	20	paper_proposes	20	phrase_based

2013		2014	
Rank	Term	Rank	Term
1	machine_translation	1	machine_translation
2	state_art	2	relation_extraction
3	social_media	3	statistical_machine
4	statistical_machine	4	neural_network
5	natural_language	5	dependency_parsing
6	part_speech	6	cross_lingual
7	coreference_resolution	7	recurrent_neural
8	significant_improvements	8	semi_supervised
9	latent_variable	9	sentence_level
10	question_answering	10	word_segmentation
11	phrase_based	11	distributional_semantics
12	relation_extraction	12	document_summarization
13	large_scale	13	distant_supervision
14	cross_lingual	14	domain_adaptation
15	semi_supervised	15	question_answering
16	sense_disambiguation	16	empirical_study
17	pos_tagging	17	similarity_contextual
18	experimental_results	18	word_sense
19	distant_supervision	19	neural_networks
20	results_show	20	weakly_supervised

Appendix 4: Top 10 nodes with high degree per year

2007			2008			2009	
Rank	Author		Rank	Term		Rank	Term
1	Chris Dyer		1	Ming Zhou		1	Ming Zhou
2	Wade Shen		2	Ting Liu		2	Heng Ji
3	Matthew Purver		3	Qun Liu		3	Mary Harper
4	Chris Callison-Burch		4	Sheng Li		4	Noah Smith
5	Sebastian Varges		5	Chin-Yew Lin		5	Yang Liu
6	Yao Meng		6	Xiaofeng Yang		6	Sibel Yaman
7	Fuliang Weng		7	Alessandro Moschitti		7	Tiejun Zhao
8	Badri Raghunathan		8	Mark Dredze		8	Chris Dyer
9	Harry Bratt		9	Min Zhang		9	Haizhou Li
10	Zhaoxia Zhang		10	Vladimir Edelman		10	Qun Liu

2010			2011			2012	
Rank	Term		Rank	Term		Rank	Term
1	Chris Dyer		1	Chris Dyer		1	Ming Zhou
2	Dan Klein		2	David Chiang		2	Qun Liu
3	Qun Liu		3	Ming Zhou		3	Chris Callison-Burch
4	Chris Callison-Burch		4	Yang Liu		4	Chris Dyer
5	Chris Quirk		5	Dipanjan Das		5	Yejin Choi
6	Min Zhang		6	Noah Smith		6	Chris Quirk
7	Ting Liu		7	Brian Roark		7	Makr Dredze
8	Mark Dredze		8	Dan Klein		8	Regina Barzlay
9	David Chiang		9	Eduard Hovy		9	Karl Stratos
10	John DeNero		10	Joakim Nivre		10	Noah Smith

2013			2014	
Rank	Term		Rank	Term
1	Chris Dyer		1	Ming Zhou
2	Chris Callison-Burch		2	Heng Ji
3	Ming Zhou		3	Kevin Knight
4	Mark Dredze		4	Chris Callison-Burch
5	Noah Smith		5	Chris Dyer
6	Tiejun Zhao		6	Martha Palmer
7	Pushpak Bhattacharyya		7	Noah Smith
8	Benjamin Van Durme		8	Yoav Goldberg
9	Joakim Nivre		9	Steven Bethard
10	Kevin Knight		10	Ting Liu

Appendix 5: Topic and institution distribution in MT network of 2013

The 2013 publications focused on machine translation were selected for this case study. The collaboration network hubs for the largest connected component were manually tagged for main topic categories and institutions, based on the titles and content of the publications. The main hubs appear to be in the U.S. (East Coast) and China.

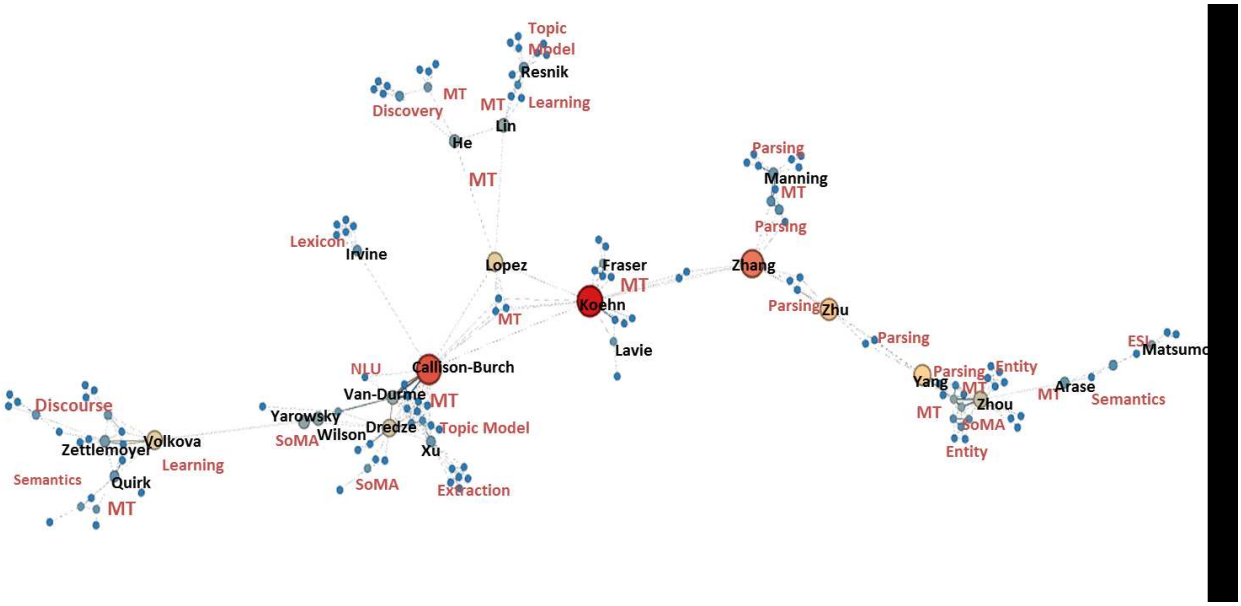

Figure 19 - Topic categories in the MT network of 2013

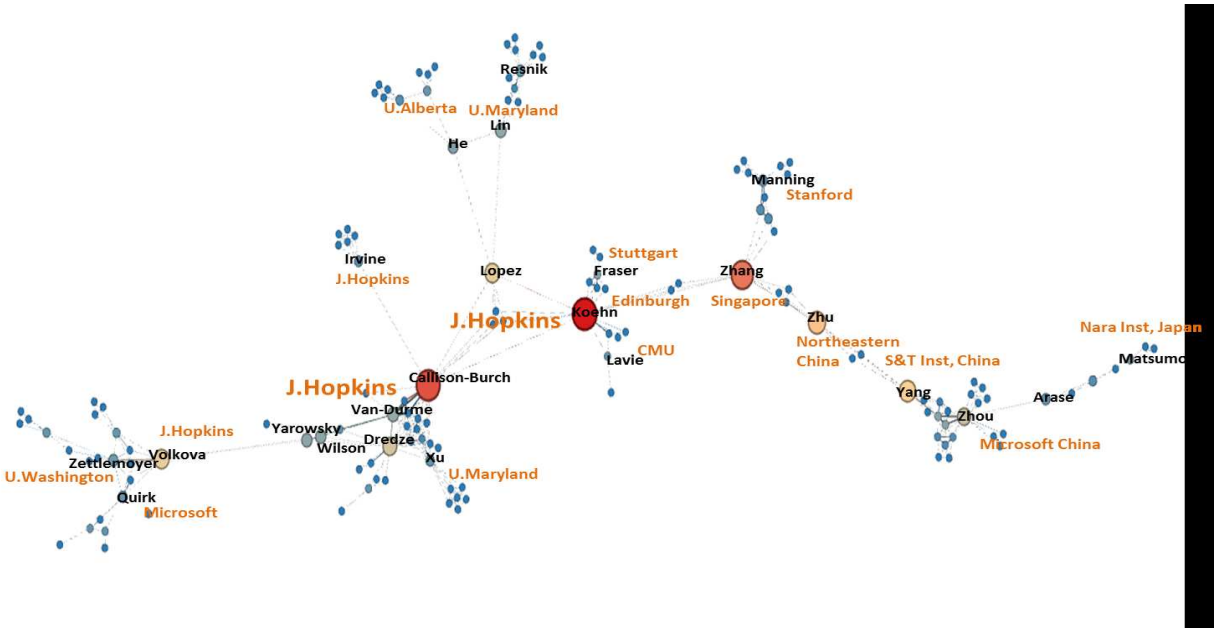

Figure 20 - Main institutions in the MT network of 2013

The second largest component in the 2013 network focuses on enabling technologies, including parsing. This cluster also contains a subcluster towards the bottom that focuses on Arabic NLP, semantics and sentiment analysis.

Figure 21 – Main topics (left) and institutions (right) in the parsing network of 2013