The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence
Symbiotic Cognitive Systems: Technical Report WS-16-14

Symbiotic Cognitive Computing through Iteratively Supervised Lexicon Induction

Alfredo Alba, Clemens Drews, Daniel Gruhl, Neal Lewis,
Pablo N. Mendes, Meenakshi Nagarajan, Steve Welch
IBM Research Almaden
San Jose, California 95120, USA

Anni Coden
IBM TJ Watson Research
Yorktown Heights, NY 10598, USA

Ashequl Qadir[*]
University of Utah
Salt Lake City, Utah 8411, USA

Abstract

In this paper we approach a subset of semantic analysis tasks through a symbiotic cognitive computing approach – the user and the system learn from each other and accomplish the tasks better than they would do on their own. Our approach starts with a domain expert building a simplified domain model (e.g. semantic lexicons) and annotating documents with that model. The system helps the user by allowing them to obtain quicker results, and by leading them to refine their understanding of the domain. Meanwhile, through the feedback from the user, the system adapts more quickly and produces more accurate results. We believe this virtuous cycle is key for building next generation high quality semantic analysis systems.

We present some preliminary findings and discuss our results on four aspects of this virtuous cycle, namely: the intrinsic incompleteness of semantic models, the need for a human in the loop, the benefits of a computer in the loop and finally the overall improvements offered by the human-computer interaction in the process.

Introduction

Text understanding in a domain requires the ability to recognize domain specific concepts within context; these can be nouns (things in the domain), verbs (actions unique to the domain) and adjectives/adverbs (modifiers that have particular meaning in the domain) as well as other phrases with special meaning. This domain semantics analysis requires the creation of relevant semantic assets (lexicons, taxonomies, annotated text, etc.) and algorithms that are able to intelligently use those assets.

Traditional approaches for building semantic analysis systems fall into the following categories:

- reliance on existing semantic assets (standard vocabularies, ontologies and lexicons),
- manually annotated sentences generated before a system is developed,
- post-development feedback after system development.

We argue that these approaches miss out on the opportunity to improve both the machine and the human efficiency in the overall task. With an iterative human and computer in the loop approach, the computer model training can be done faster and the human understanding of the task can be improved. A new class of systems is required; one that leverages human input efficiently, early and continuously in the semantic analysis process. This is the class of symbiotic cognitive computing systems that is the focus of this work.

We describe in this paper one example of such system. We call it *Iteratively Supervised Lexicon Induction* (***ISLI***). We organize our discussion in terms of four aspects:

- accepted mature standard semantic assets are seldom complete or constantly evolving;
- including a human in the loop when creating semantic assets leads to more complete and accurate results;
- using a cognitive system to help a human create semantic assets improves human understanding of the task;
- a symbiotic cognitive computing system leads to higher quality, lower cost solutions.

We show experimental validation of the hypothesis that human involvement at each step improves the overall results, as compared to the traditional approach of leaving the adjudication to the end user (or developer) at the end of the (automatic) induction process. This is most evident in high-value domains, where errors in a lexicon can be very costly. While the evidence we present is not unequivocal, it supports the human in the loop hypothesis and suggests several avenues for additional exploration.

Framework

Cognitive computing systems often require language understanding as a way to naturally interact with users. Language processing systems are often organized in a pipeline,

[*] Work performed during internship at IBM Research Almaden.

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

where upstream tasks send their outputs to be used as input in downstream tasks. Most text analytic systems are reliant on upstream semantic tagging of the text. Without accurate complete tagging of possible meanings for words and phrases is is unlikely such systems can derive the correct results. This applies to a wide variety of systems and has been shown by successful implementations of question answering (Ferrucci et al. 2010), entity extraction (Mendes et al. 2011) and sentiment analysis (Coden et al. 2014).

In this paper we will focus the discussion around a pipeline where the user starts with a collection of terms of interest (i.e. a lexicon) that will help them conduct text analytics on a given corpus. The initial list of terms is expanded using a system we call *Iteratively Supervised Lexicon Induction* (***ISLI***). The ***ISLI*** engine seeks to co-develop concept phrases and the patterns around them in a semi-supervised process. It first takes a series of seed phrases that describe the concept of interest and examines the corpus to generate all context patterns that appear around them. For example, if "fox" was a seed phrase, the occurrence of "the quick brown fox jumped over" would generate patterns of "brown *", "* jumped", "brown * jumped", etc. up to six words on either side. This set of patterns (usually several million) is then applied to the corpus (or a sample of it if the corpus is very large). For each pattern the phrases that replace the * (up to six words) are noted. The pattern is scored for support (how many different phrases fill it) and confidence (of the phrases that fill it, how many are known good phrases). Patterns that are good enough (support and confidence above a threshold) are kept, and for each candidate phrase the prevalence (number of good patterns it appears in) is computed. The candidate phrases are presented in this order and good ones are added to the seed phrase list by a subject matter expert. The process is then repeated. For more algorithmic detail (Coden et al. 2012)(Qadir et al. 2015) or the impact of ***ISLI*** in downstream applications (Coden et al. 2014), please refer to previous work.

Evaluation

Standard lexicons always need to be expanded

To objectively evaluate the impact of expanding a lexicon with iterative supervision, we conduct an experiment with sentiment lexicons widely used in the sentiment analysis literature.

We obtained the SemEval 2014 task 4 (Pontiki and Bakagianni 2013) dataset collected from restaurant reviews. It contains 2,179 positive sentiment sentences (3,779 unique words), as well as 839 negative sentiment sentences (2,376 unique words). For comparison, we obtained 3 sentiment lexicons that are widely used for sentiment analysis purposes – e.g. sarcasm detection (Riloff et al. 2013). These lexicons are:

- ***Liu et al.*** (Liu, Hu, and Cheng 2005) – containing 2,007 positive sentiment words and 4,783 negative sentiment words;
- ***ANEW*** (Nielsen 2011) – containining 1,598 positive sentiment words and 879 negative sentiment words; and
- ***MPQA*** (Wilson et al. 2005) – containing 2,718 positive sentiment words and 4,910 negative sentiment words.

We counted the number of hand-labeled positive sentences in the gold standard corpus containing a positive term, as well as the number of hand-labeled negative sentences containing negative terms, for each of the lexicons. We found that, in this corpus, the ***MPQA*** lexicon has the highest coverage of sentiment expressions annotations, when compared to ***ANEW*** and ***Liu et al.*** (Table 1).

Table 1 shows that a significant portion of the sentences do not match any sentiment words in these *de facto* standard lexicons. We hypothesized that there are idioms or other non-standard expressions that could be discovered to increase the lexicon coverage over this dataset. Thus, we used the ***MPQA*** corpus as seed, and expanded its entries through ***ISLI*** on all available corpora. In fact, after repeating the experiment with the extended dictionary (***MPQA +ISLI***), we observed an additional 5% and 1% coverage for the positive and negative sentiment sentences respectively compared to the ***MPQA*** lexicon.

Human-in-the-loop outperforms automatic expansion

We conducted experiments on the MUC4 corpus (Sundheim 1992), one of the most widely used corpora to evaluate semantic lexicon creation (Riloff and Shepherd 1997; Roark and Charniak 2000; Riloff and Jones 1999; Thelen and Riloff 2002). We used ***ISLI*** over MUC4's 1700 text documents and compared the results to a gold standard[1] released in 2009. It includes seven categories: BUILDING, EVENT, HUMAN, LOCATION, TIME, VEHICLE, and WEAPON. The corpus is topic-focused and the terms are limited to head nouns – one of the authors manually labeled every head noun in the corpus that was found by their extraction patterns. As a consequence, only 5.36% (210/3911) of gold standard terms have more than one token. We used MetaBoot (Riloff and Jones 1999) as a reference system to allow us to compare the effect of iteratively supervising a system, in contrast to automatically inducing a lexicon and curating it at the end.

Precision and recall Table 2 reports the precision and recall of ***ISLI***, showed side by side with MetaBoot's at the point where both systems have acquired 1,000 terms. The precision is defined as the number of terms in the semantic lexicon that are also in the gold standard, while the recall is defined as the portion of the gold standard that has already been added to the semantic lexicon.

For all types, ***ISLI*** has achieved higher recall than MetaBoot. We believe that this is due to two reasons. First, ***ISLI*** focuses on using multiple patterns of average accuracy, rather than focusing on only the best patterns for each lexicon. Second, the iterative supervision allows ***ISLI*** to keep only good terms at each iteration, and use those to derive better patterns early on in the process while keeping spurious patterns from dominating.

For MetaBoot, the numbers shown are the result of automatically collecting the top 5 terms in each iteration for 200

[1] Available from: http://www.cs.utah.edu/~jtabet/basilisk/

	Positive Sentences with Match	**Negative Sentences with Match**
Liu et al. (Liu, Hu, and Cheng 2005) Lexicon	73.34%	42.55%
ANEW (Nielsen 2011) Lexicon	65.72%	30.63%
MPQA (Wilson et al. 2005) Lexicon	77.97%	41.70%
MPQA + ISLI	82.93%	51.97%

Table 1: Coverage comparison of different lexicons on a sentiment dataset

	Precision	**PBC**	**Recall**	
	MetaBoot	***ISLI***	MetaBoot	***ISLI***
Building	0.043	0.096	0.228	0.511
Event	0.190	0.209	0.379	0.417
Human	0.278	0.554	0.149	0.298
Location	0.310	0.313	0.305	0.307
Time	0.026	0.050	0.232	0.464
Vehicle	-	0.043	-	0.483
Weapon	0.033	0.078	0.224	0.531

Table 2: Precision and Recall at 1000 terms for ***ISLI*** and for MetaBoot as reported in (Thelen and Riloff 2002)

iterations. For ***ISLI***, it is the result of showing an average of 100 terms per iteration for 10 iterations. Recall that in ***ISLI*** the user is involved in curating the term candidates. Hence, the measured precision after each iteration (as defined by related work) would be 100%, since it reflects the expert's understanding of the gold standard at that point in time. Therefore we show the Precision Before Curation (PBC) in the Precision column in Table 2 for ***ISLI***: the percentage of term candidates that were sensible – i.e., contained a gold standard term. This percentage was also higher than the proportion of terms found by MetaBoot that contained head nouns from the gold standard.

This confirms our hypothesis that involving the human in each iteration produces better results, increasing the final recall and reducing the amount of effort in lexicon cleanup (higher precision).

Missing, generic, specific and implied terms Contradicting the expectation that the gold standard would contain an exhaustive set of terms in each category, our evaluation through ***ISLI*** enabled the discovery of a number of new terms. We initialized ***ISLI*** with the full set of terms provided in the MUC4 gold standard, and ran one iteration over the full MUC4 corpus.

We organized the newly found terms into four groups: missing, misspellings, specific and implied. **Missing** terms were those that, from our understanding, should have been present in the gold standard because they seem to follow the same general trend as the terms already present there. For example, in the Building category, the term 'sheraton' was present, but not 'hilton'; Weapon contained 'ak-47' but not 'uzi' and Vehicle contained 'trucks' but not 'jitneys'. **Misspellings**. Although several terms were present in the gold standard with alternative spellings, the list of spelling variants or misspellings was not complete. Therefore we collected those under the category Misspelling. These include 'discoteque' vs 'discotheque', 'libersation' vs 'liberation', etc. **Specific** terms were those that illustrate a particular aspect that we consider important in the context of the semantic lexicon, but that were not annotated. In the category Human, the term 'bustillos' was present, but not 'juan rafael bustillos', while in Vehicle, there was 'helicopter' but not 'hughes 500 helicopter'. The difference in these terms might be the difference between two people in the same family, or between a military and a civilian vehicle, and therefore may be non-negligible depending on the specificity required by the task. **Implied** terms were those that refer to an entity of a given semantic type, as an indirect reference, or a generic term. For example, we extracted 'unidentified speaker' as a term in the Human category, and 'john f. kennedy' (the carrier) as Vehicle. Table 3 shows one out of many examples that were found in each of these groups along with the percentage of new examples that were found after 2 iterations of ***ISLI***.

Computer-in-the-loop improves human understanding

We also experimented with lexicons with very different characteristics with respect to the types of terms that are allowed in the lexicon. Through the human-computer interaction in building these lexicons, we could observe that the human experts were learning and adapting their understanding of the lexicons throughout the expansion process.

The lexicon Affinity contains terms such as 'love' and 'adore' that indicate someone has a personal preference for something. Phrases such as 'have an obsession with' surfaced and forced the user to refine their membership criteria to delimit boundaries between affinity and pathological attachment. Polarity contains terms people use to review something. While affinity such as 'gr8' may be included, it also includes more objective terms such as 'reasonably priced' and 'convenient', as well as idioms such as 'like a charm'. Voters contains terms used in the congressional record when a politician is referring to either all voters or a block of voters, e.g., 'American people', 'working families', 'the public'. One surprise here is the list is relatively short; congressional rhetoric appears to be fairly stylized. Symptoms contains issues a patient conveys to a doctor that cannot be objectively verified. For example, 'pain' or 'inability to stay asleep' or 'loss of sense of smell'. Since symptoms are classically reported "in the patients' own words" they are one of the most variable parts of the clinical record, and show substantial variation depending on the culture and background of the patients seen. Seven Words was created starting with George Carlin's "Seven Words you can't say on TV". The system found terms containing ⋆s as a way to mask their original spelling, slang variations, etc.

Type	Missing	Specific	General/Implicit	Growth
Building	hilton	presidential house	hotel lobby	63.30%
Event	retreat	guerrilla attack	wave of violence	8.58%
Human	bernard	juan rafael bustillos	unidentified speaker	7.22%
Location	la paz	san jose (costa rica)	salvadoran capital	21.12%
Time	may	25th of april	saturday night	110.71%
Vehicle	jitneys	hughes 500 helicopter	john f. kennedy (aircraft carrier)	53.93%
Weapon	uzi	dragunov rifle	guerrilla	29.93%

Table 3: Examples of the types of terms discovered through ***ISLI***, along with the number of correct terms added to the gold standard.

This rapidly develops a list of terms that might not be appropriate in polite conversation. It highlights users creativity in evading automatic profanity blockers and its importance in understanding strong opinions about a product or event. WORRY contains terms used by people when they are worried about something. These can be useful signposts for spotting emails sent to a contact address that might warrant a personal response or when examining a large corpus of customer concerns with an important issue (e.g., getting a car loan, seeing a physician about a sensitive subject, finding a lawyer), or when helping a service provider know what to reassure potential customers of in their advertisements (e.g., "bad credit is not a problem", "confidential and discreet", "we're on your side").

Human-machine partnership leads to lower costs and higher quality

Another aspect of a symbiotic Human-Machine partnership is allowing for more productive use of expert time for specialized tasks. Cognitive systems can assist subject matter experts in understanding their own data and problem domains while decreasing time to delivery analytic models.

We analyzed the efficiency of human-generated annotations (without ***ISLI***) in comparison with extended pre-annotation (followed by human corrections and completion). A Medical Doctor and an assistant were tasked to create 2 ground truths that identified several entity types in text: Genes, Genetic Variants, Cancers, Cancer Treatments, and Treatment Responses. The first ground truth (#1) consisted of 10 sections from 3 medical journal articles and was pre-annotated with standard medical lexicons. The second ground truth (#2) consisted of 150 sections from 30 journal articles and was pre-annotated with medical lexicons that were extended by the MD and assistant using ***ISLI*** trained on various medical journal articles and abstracts.

Ground truth #1 manual annotation time took approximately 3 person hours ($\sim$1 article / hour), while #2 took approximately 5.5 person hours ($\sim$6 articles / hour). The medical lexicon extension time was approximately 2 person hours to increase the lexicon size by 21%. Overall, annotation time decreased by 6x overall, or 4x if lexicon extension time is included.

Named Entity Recognition (NER) models were created from each ground truth, with 70/20/10 train/test/blind dataset separation. The model without pre-annotation extensions had an F1 of .68, while models with extensions had .77, resulting in a 9% increase in NER performance.

This shows how the human-machine partnership significantly decreased the required time investment by subject matter experts (and thus, cost) while increasing the accuracy of analytics.

Conclusion

We have presented a symbiotic cognitive computing approach for semantic asset creation. We focused on lexicon creation and its impact on common downstream tasks such as text analytics, sentiment analysis and named entity recognition.

We argued that subject matter experts should be involved in the lexicon creation process, since the exploration of examples help the users make important decisions on their desired term specificity. We show that when using the MUC4 gold standard as seeds to our system, we are able to significantly extend its lexicon with terms that although present in the corpus were absent from the gold standard. Besides obvious omissions from the gold standard we highlighted interesting kinds of mentions that our system is able to discover including terms expressing significant specific differences, as well as generic and implied references that were missing from the gold standard. This highlighted human learning from the system.

We plan to expand on our experiments and report on the efficacy of applying ***ISLI*** to multiple languages, different types of discourse, and domains of knowledge, further exploring the complementary nature of cognitive processing done by humans and machines for semantic analyses.

Acknowledgements

Author names listed in alphabetical order by last name and grouped by affiliation.

References

Coden, A.; Gruhl, D.; Lewis, N.; Tanenblatt, M. A.; and Terdiman, J. 2012. Spot the drug! an unsupervised pattern matching method to extract drug names from very large clinical corpora. In *HISB*, 33–39.

Coden, A.; Gruhl, D.; Lewis, N.; Mendes, P. N.; Nagarajan, M.; Ramakrishnan, C.; and Welch, S. 2014. Semantic lexicon expansion for concept-based aspect-aware sentiment analysis. In *Semantic Web Evaluation Challenge - SemWebEval 2014 at ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*, 34–40.

Ferrucci, D. A.; Brown, E. W.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A.; Lally, A.; Murdock, J. W.; Nyberg, E.; Prager, J. M.; Schlaefer, N.; and Welty, C. A. 2010. Building watson: An overview of the deepqa project. *AI Magazine* 31(3):59–79.

Liu, B.; Hu, M.; and Cheng, J. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *WWW '05*, WWW '05, 342–351. New York, NY, USA: ACM.

Mendes, P. N.; Jakob, M.; García-Silva, A.; and Bizer, C. 2011. DBpedia Spotlight: shedding light on the web of documents. In *I-SEMANTICS 2011*, 1–8.

Nielsen, F. A. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. In *#MSM2011*.

Pontiki, M., and Bakagianni, J. 2013. Semeval-2014 absa restaurant reviews - train data.

Qadir, A.; Mendes, P. N.; Gruhl, D.; and Lewis, N. 2015. Semantic lexicon induction from twitter with pattern relatedness and flexible term length. In *AAAI'15*, 2432–2439.

Riloff, E., and Jones, R. 1999. Learning Dictionaries for Information Extraction by Multi-level Bootstrapping. In *AAAI '99/IAAI '99*, 474–479.

Riloff, E., and Shepherd, J. 1997. A corpus-based approach for building semantic lexicons. In *Second Conference on Empirical Methods in Natural Language Processing*, 117–124.

Riloff, E.; Qadir, A.; Surve, P.; De Silva, L.; Gilbert, N.; and Huang, R. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP'13*, 704–714. Seattle, Washington, USA: ACL.

Roark, B., and Charniak, E. 2000. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. *CoRR* cs.CL/0008026.

Sundheim, B. M. 1992. Overview of the fourth message understanding evaluation and conference. In *MUC'92*, MUC4 '92, 3–21. Stroudsburg, PA, USA: ACL.

Thelen, M., and Riloff, E. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *EMNLP '02*, EMNLP '02, 214–221.

Wilson, T.; Hoffmann, P.; Somasundaran, S.; Kessler, J.; Wiebe, J.; Choi, Y.; Cardie, C.; Riloff, E.; and Patwardhan, S. 2005. Opinionfinder: A system for subjectivity analysis. In *EMNLP'05 Demos*, 34–35. Vancouver, BC, Canada: ACL.