Automatic Text Understanding of Content and Text Quality

Ani Nenkova
University of Pennsylvania

Reading involves two rather different kinds of semantic processing. One is related to understanding what information is conveyed in the text and the other to appreciating the style of the text, how well or poorly it is written. For people, text content and stylistic quality are inextricably linked. For machines, robust understanding of written material has become feasible in many contexts but text quality has been out of reach so far. The mismatch matters a great deal because people rely on machines to locate and navigate information sources and increasingly read machine generated text, for example as machine translations or text summaries.

In this presentation I will discuss some of the simple and elegant intuitions which have enabled semantic processing in machines, as well as some of the emerging directions in text quality assessment.

1. Text semantics (meaning)

1.1 Reading and understanding the Web

A single insight about language semantics has led to successes in a variety of automatic text understanding tasks. Words tend to appear in specific contexts and these contexts convey rich information about the type of the word, its meaning and connotation [Harris, 1968]. Computers can learn much semantic information without human supervision, simply by collecting statistics of (hundreds of) thousands of texts.

The context of a target word, consisting of other phrases or words that occur nearby in texts more often then expected by chance, is accumulated over large text collections. For example the word `tea` may be characterized by the context `[drink:60, green:55, milk:40, sip:30, enjoy:10, …]`. Each entry shows a word that appeared five words before or after `tea`, and the number of times the pair was seen in a large text collection. Taking just the number of occurrence of context words makes the representation even more convenient, because various standard (geometric) approaches exist for comparing the distance between numeric vectors. In this manner, a machine can compute the similarity between any two words.

Here is an example from Pantel and Lin [2002] of the 15 words most similar to`wine` computed by this approach.

> **`Wine`**`: beer, white wine, red wine, Chardonnay, champagne, fruit, food, coffee, juice, Cabernet, cognac, vinegar, Pinot noir, milk, vodka,…`

The list may not look immediately useful but is certainly impressive if one considers how little similarity there is in the sequence of letters wine, beer, Chardonnay.

Building upon these representations, it has become possible to automatically discover words with multiple senses by clustering words similar to them (plant: (plant, factory, facility, refinery) (shrub, ground cover, perennial, bulb)), find synonyms and antonyms. To aid analysis of customer reviews, researchers at Google developed a large lexicon of almost 200,000 positive and negative words and phrases, identified through their similarity to a handful of predefined positive or negative words such as excellent, amazing, bad, horrible. Among the positive phrases in the automatically constructed lexicon were cute, fabulous, top of the line, melt in your mouth; negative examples included subpar, crappy, out of touch, sick to my stomach [Velikovich et al, 2010].

Another line of research in semantic processing exploits the stable meaning of some contexts. For example patterns like ``X such as Y", if occurring often in texts, is very likely an indicator that Y is a kind of X, i.e. "Red wines such as Cabernet and Pinot noir…". Similarly a phrase like "The mayor of X" is a good indicator that X is a city. NELL (Never Ending Language Learning, http://rtw.ml.cmu.edu/rtw/) is a system that constantly learns unary and binary predicates, corresponding to categories and relations such as isCity(Philadelphia) and playsInstrument(George_Harrison, guitar). The learning of each type of fact starts with minimal supervision in the form of several examples of category instances or entities between which a relation holds, given by the researchers. Then the system starts an infinite loop in which it finds web pages that contain the examples, finds phrase patterns that typically occur with the examples, selects the best patterns which indicate the predicate with high probability, then applies the patterns to new texts to discover more instances for which the predicate is true. Different flavors of this approach to machine understanding have been developed to help search and question answering [Etzioni et al, 2008, Pasca et al 2006].

1.2 Reading and understanding a text

In the semantic processing I have discussed so far, the computer reads numerous textual documents with the objective to learn representations of words, come up with lexicon of phrases with positive or negative connotation, or learn category instances and relations. A more difficult task for a computer is to understand a specific text.

Much traditional research related to computer processing of a single text has relied on supervised techniques. Researchers invested effort to prepare collections in which human annotators marked positive and negative examples of a semantic distinction of interest. For example they could mark the different senses of a word, the part of speech of words, or would mark that Roger Federer is a person, Bulgaria is a country. Then features describing the context of the categories of interest would be extracted from the text, and a statistical classifier would use the positive and negative examples to combine the features and predict the same type of information on unseen text. More recently it has become clear that the unsupervised approach in which computers accumulate knowledge/statistics from large amounts and text and the supervised approach can be combined effectively and result in better systems for semantic processing.

When reading a specific text, computers also need to resolve what entity in the document are referred to by pronouns such as “he/his”, “she/her”, “it/its”. Systems are far from perfect but are getting better at this task. Usually pronouns appear in the text nearby noun phrases, i.e. “The *professor* prepared *his* lecture” but in other situations gender and number information is necessary to correctly resolve the pronoun, as in “*John* told Mary *he* had booked the trip”. Machines can rather accurately learn the likely gender of names and nouns, again by reading large volumes of text, and collecting statistics of co-occurrence. Statistics about the co-occurrence of a pronoun of a given gender and the immediately preceding noun or honorifics and names (Mr. John Black, Mrs. Mary White), collected over thousands of documents, give surprisingly good guesses about the likely gender of nouns [Bergsma, 2005].

2. Text quality (style)

Automatic assessment of text quality, or style, is a far more difficult task compared to acquisition of semantics, or at least considerably less researched. Much of the effort in my lab has been focused on developing models of text quality. I will discuss two successful endeavors: prediction of general and specific sentences and automatic assessment of sentence fluency in machine translation and summary coherence in text summarization.

A well-written text contains a balanced mix of general overview statements and specific detailed sentences. If a text contains too many general sentences it will be perceived as insufficiently informative, and too much specificity can be confusing for the reader.

To train a classifier, we exploit a resource of 1 million words of Wall Street Journal text with discourse annotations [Louis and Nenkova, 2011]. The discourse annotations, among other things, specify the way in which two adjacent sentences in the text are related. There could be an implicit comparison between two statements (John is always punctual. Mary often arrives late.), or a contingency (causal) relation (I hurt my foot. I cannot go dancing tonight.), or temporal relations.

One of the discourse relations annotated in the corpus is instantiation. It holds between two adjacent sentences where the second gives a specific example of information mentioned in the first, as in “He is very smart. He solved the problem in five minutes”. We considered that the first sentence is general while the second is specific in all instances of instantiation relation. We computed a number of features which according to our intuition would distinguish between the two categories. We expected that the presence of opinion or evaluative statements would characterize the general sentences, as well as unusual use of language that would later be interpreted or clarified in a specific sentence. Among the features were

- the length of the sentence
- the number of opinion or subjective words, derived from an existing dictionaries

- the specificity of words in the sentences, derived from corpus statistics as the fraction of documents in a one year of New York Times articles that contain the word. The fewer documents contain the word, the more specific it is.
- mentions of numbers and people, companies and geographical locations; such mentions are detected automatically.
- syntactic features related to adjectives, adverbs, verbs and prepositions
- probabilities of sequences of one, two or three consecutive words computed over one year of New York Times articles.

A logistic regression classifier, trained on around 2,800 examples of general and specific sentences from instantiation relations, learned to predict the distinction incredibly well. On a completely independent set of news articles, five different people were asked to mark each sentence as general or specific. For sentences in which all five annotators agreed about the class, the classifier can predict the correct class with 95% accuracy. For examples on which only four out of the five annotators agreed, the accuracy is 85%. For all examples, which included sentences for which people found it hard to classify in terms of general and specific, the accuracy of prediction was 75%. Moreover, the confidence of the classifier turned out to be highly correlated with annotator agreement, so it was possible to identify which sentences will not fit squarely into one of the classes. The degree of specificity of a sentence given by the classifier gives an accurate indication of how a sentence will be perceived by people.

Applying the general/specific classifier to samples of automatic and human summaries of clusters of news articles has demonstrated that machine summaries are overly specific and has indicated ways for improving system performance [Louis and Nenkova, 2011].

Word co-occurrence statistics and subjective language have also been successful in automatically distinguishing implicit comparison, contingency and temporal discourse relations [Pitler et al, 2009]. Identification of such relations is not only necessary for semantic processing of text, it is also required for robust assessment of text quality [Pitler and Nenkova, 2008]. Finally, statistics on types, length and distance between verb, noun and prepositional phrases, as well as probabilities of occurrence and co-occurrence of words are highly predictive of the perceived quality of summaries [Nenkova et al, 2010].

[Bergsma, 2005] Shane Bergsma. 2005. Automatic Acquisition of Gender Information for Anaphora Resolution, *Proceedings of the 18th Conference of the Canadian Society for Computational Studies of Intelligence (Canadian AI'2005*), 342—353.

[Etzioni et al, 2008] Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web.*Commun. ACM* 51, 12 (December 2008), 68-74.

[Harris, 1968] Zelig S. Harris. *Mathematical Structures of Language*. Wiley, New York, 1968

[Louis and Nenkova, 2011a] Automatic identification of general and specific sentences by leveraging discourse annotations, In *Proceedings of the International Joint Conference in Natural Language Processing*, (to appear)

[Louis and Nenkova, 2011b] Annie Louis and Ani Nenkova. 2011. Text Specificity and Impact on Quality of News Summaries, In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, 34—42

[Nenkova et al, 2010] Ani Nenkova, Jieun Chae, Annie Louis, and Emily Pitler. 2010. Structural features for predicting the linguistic quality of text: applications to machine translation, automatic summarization and human-authored text. In *Empirical methods in natural language generation*, Emiel Krahmer and Mariet Theune (Eds.). Lecture Notes In Artificial Intelligence, Vol. 5790. Springer-Verlag, Berlin, Heidelberg 222-241.

[Pantel and Lin, 2002] Patrick Pantel and Dekang Lin. 2002. Discovering Word Senses from Text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2002*. pp. 613-619.

[Pasca et al, 2006] Marius Pasca, Dekang Lin, Jeffrey Bigham, Andrei Lifchits, and Alpa Jain. 2006. Names and similarities on the web: fact extraction in the fast lane. In*Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (ACL-44).

[Pitler et al , 2009] Emily Pitler, Annie Louis and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text, *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 683-691.

[Pitler and Nenkova, 2008] Emily Pitler and Ani Nenkova. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality, In*Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP)

[Velikovich et al, 2010] Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of web-derived polarity lexicons. In*Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (HLT '10). 777-785.