Mona Diab

Research Statement

Ambiguity is an inherent characteristic of natural language, permeating its various levels of representation. From a human language processing perspective, ambiguity is not a severe problem. However, from a machine processing perspective, the story is quite different. As a computational linguist, my research agenda spans several levels and perspectives on semantic ambiguity resolution in natural language. I started by exploring issues of sense ambiguity resolution in my thesis work. Currently in my postdoctoral work, I am involved in syntactic parsing and automatic semantic role labeling of predicate arguments pertaining to verbs and nominalized nouns in Arabic. I also examine issues in homonymy and polysemy distinctions from a cross-linguistic perspective and the role such a distinction can play in the organization of taxonomies and ontologies. Moreover, I am interested in issues related to metaphor detection and interpretation. In general, my research theme revolves around exploring the extent to which cross-linguistic similarities and divergences can serve as a source of evidence for ambiguity resolution in natural language processing systems.

Dissertation

My dissertation work focused on Word Sense Disambiguation (WSD) from a multilingual perspective. WSD occupied center stage in the early work on computational linguistics. With the on-going surge in machinery allowing for the development of sophisticated techniques and algorithms, WSD is experiencing a revival of interest especially with the belief that it has the potential of improving several central tasks in natural language processing. For the purposes of my thesis, I adopt the definition whereby WSD is the process of resolving the meaning of a word unambiguously in a given natural language context by marking it with explicit sense labels from a predefined tag set.

What constitutes a sense is a subject of great debate. An appealing perspective aims at defining senses in terms of their multilingual correspondences, but to date was not given any practical demonstration. My thesis constitutes an empirical validation of such a notion. In the scope of my dissertation work, word meaning is characterized in terms of its cross-linguistic correspondences. The intuitive idea is that word meaning or word sense is quantifiable in as much as it is uniquely translated in some language or set of languages.

Consequently, I address the problem of WSD from a multilingual perspective; I expand the notion of context to encompass multilingual evidence. I devise a new approach – SALAAM -- to resolve word sense ambiguity in natural language, using a source of information that was never exploited on a large scale for WSD before. SALAAM is an unsupervised approach since it does not rely on the existence of sense annotated data. SALAAM annotates large amounts of texts in translation (a parallel corpus) using only a sense inventory for one of the languages of the parallel corpus. SALAAM results in large amounts of sense annotated data in both languages of the parallel corpus, simultaneously, even if one of the languages lacks a sense inventory. Accordingly, SALAAM provides a means of bootstrapping resources for scarcely represented languages. SALAAM's performance outranks all other unsupervised approaches based on monolingual data when compared on the same data set. The interesting aspect of this work lies in the fact that it uses an orthogonal source of evidence, thereby allowing ample room for extension.

The automatic unsupervised tagged data produced by SALAAM is further utilized to bootstrap a machine learning based unsupervised learning WSD system with the intent of addressing the resources acquisition ---training data --- bottleneck for supervised methods. Essentially, SALAAM is extended as an unsupervised approach for WSD within a learning framework; in many of the cases of the words disambiguated, SALAAM coupled with the machine learning system is competitive with the performance of a canonical supervised WSD system that relies on human tagged data for training. I am able to successfully identify factors that affect the performance of the WSD system when trained using noisy input. This allows for predicting which data items do not need manual annotations (may yield equal performance if trained with SALAAM tagged data). Such information leads to a reduction of 41% in the manual annotation labor for supervised WSD systems.

Realizing the fundamental role of similarity for SALAAM, I investigate different dimensions of semantic similarity for verbs since they are relatively more complex entities than nouns. I design a human judgment experiment to obtain human ratings on verbs' semantic similarity. The obtained human ratings are cast as a reference point for comparing different automated similarity measures that crucially rely on various sources of information: distributional information, syntagmatic information, and paradigmatic information. Finally, a cognitively salient model integrating human judgments in SALAAM is proposed as a means of improving its performance on sense disambiguation for verbs.

Current Research

I am fortunate enough to have the opportunity to work with two of the best researchers in our field, Dan Jurafsky and Chris Manning. I am involved in building an Arabic semantic parser with Dan Jurafsky; and also working on an Arabic syntactic parsing with Chris Manning. Similar to English semantic parsing, Arabic semantic parsing addresses issues of role annotations of arguments and adjuncts of clausal predicates. The approach we adopt utilizes machine learning techniques for the identification and labeling of relevant entities in a sentence. Unlike the English data however, there exist no Arabic propositional bank (propbank) for training. Therefore, part of my job is to manually create a sample Arabic propbank. I have created annotations for 1000 sentences covering the 10 most frequent verbs in the Arabic TreeBank. On the other hand, we are developing an Arabic syntactic parser based on the Stanford English factored parser. This opportunity allows me to think deeply of problems with modern standard Arabic grammar and semantic representations. Moreover, since syntactic and semantic parsers currently exist for English, and Chinese, I have the unique opportunity to explore the cross linguistic space between the three languages. Specifically, I am interested in investigating the different ways various languages split up the semantic space exemplified in their surface representations.

In the process of working on these projects, I have developed [in conjunction with Kadri Hacioglu and Dan Jurafsky] a set of basic state-of-the-art processing tools for the tokenization, POS tagging and base phrase chunking of Arabic text.

As a result of my thesis work, and my current interest and involvement in Arabic language processing, I automatically bootstrap a seed Arabic WordNet. WordNets are taxonomies of language. They are currently used as a standard lexical resource for NLP work. I evaluate the resulting automated bootstrapped WordNet structures in several human subject rating experiments with very promising performance.

Another aspect of my research investigates the extent an explicit distinction between homonymy and polysemy is automatically detectable and useful for concept organization within ontological

and taxonomical representations of natural language. This line of research is supported by interesting psycholinguistic evidence (Rodd et.al. 2000, 2001, 2002) strongly suggesting that the distinction is explicit in our mental lexicon. The intuition, in this case, is that if we make such a distinction explicit it will have a positive ripple effect on the way in which we build computational models of meaning representation, annotation and eventually evaluation.

Future Research

In relation to my thesis work, to SALAAM:

I will extend my work, which strictly deals with texts in translation, parallel corpora, to deal with texts that are related to the same genre and the same timeline but not strict translations of one another, known as comparable corpora. Though the data is more readily available, the problem becomes more challenging, since it will hinge upon issues of paraphrase detection and utterance similarity.

Moreover, I intend to extend the SALAAM approach to take monolingual evidence into consideration since such evidence is explicitly ignored in the current implementation. Such a combination of evidence, monolingual with cross-lingual, should yield better performance and prune the large search space currently utilized.

As mentioned above, I characterize the factors that predict which data items need manual annotation for a supervised WSD system. I intend to combine these different factors automatically in order to discern these cleanly SALAAM tagged items in order to reduce the manual data acquisition effort.

Finally, I am interested in extending my verb similarity study to encompass more verb pairs and more human subjects. The goal is to use the results in a realistic predictive model for SALAAM's similarity kernel, but, moreover, I would investigate the correlation between human intuitions and automated models of semantic similarity for complex grammatical categories.

With respect to general future research goals:

This summer, July-August 2005, I will be one of the senior leaders of an NSF funded workshop through the Johns Hopkins University. The subject of the workshop (presented in conjunction with Nizar Habash, & Owen Rambow from Columbia University) is the syntactic parsing of Arabic dialects. As is well known, the Arabic language is one that exhibits diaglossia between spoken dialect and written text as in Modern Standard Arabic (MSA). Many resources and tools are being constructed for MSA, while dialect suffers from the lack of resources, therefore creating the computational challenge. On the other hand, we would like to explore the extent of predictability of variation between MSA and dialect and its usefulness in tackling the issue of parsing. Hence, the crux of our workshop will be to exploit the similarities between dialect and MSA as exemplified in linguistic studies for solving the challenge of parsing Arabic dialect. In the preparation for this workshop, we will create resources and tools to process Arabic dialects – specifically Egyptian and Levantine Arabic.

In a very different stream of work, in collaboration with Jim Martin, I intend to investigate ways of detecting different types of metaphors in natural language text. We realize that figurative language permeates text and if we are to build realistic models of natural language we need to

have some means of detecting and processing metaphors as well as metonymical constructions. In fact, it has been noted that approximately 54% of the verbs in the WSJ are used metaphorically. Our approach will be corpus based and would build on insights from cross-linguistic typological studies and evidence.

My main goal in general, is to use the knowledge I acquired during my studies and my postdoctoral research to create innovative tools and methodologies that could enhance the field of computational linguistics by building on profound linguistic insights coupled with sound computational principles.