Integrated Research Plan

One of the many motivations or making machines efficient language learners is for them to be able to translate between the diverse variety of languages out there. Currently the dominant approaches to machine translation lay their foundation on pure statistical techniques that learn to translate using huge amounts of parallel data. The key reason for why they seem to work well for resource rich languages like English (and a few alike) is because they are trained on frequently occurring material and most of what people say is repetitive and predictable. But for more than 95% of languages that fall under low resource languages, such huge amounts of data might never be available. But human learners do not need such huge amounts of data to be efficient translators. So how do they do that? The answer to this is not yet fully understood but at a very basic level one would agree that humans try to process the “meaning” of the sentence before they attempt to translate it. Hence since the inception of machine translation it has been claimed that a semantic approach is required to achieve human-like translation (Weaver, 1955; Bar-Hillel, 1960). One plausible semantic based approach to translation would be to go from the source sentence to some meaning representation and then construct back a sentence in the target language from the representation. Abstract Meaning Representation (Banarescu et al., 2013) (AMR) is one such meaning representation. In this work I want to explore how we can use AMR for creating a translation system that can be trained on medium sized data and that can make use of some of the key linguistic intuitions that human language learners use. I believe that my prior experience in machine translation and AMR parsing puts me in a good starting position to explore the computational aspect of this work. For gaining the required linguistic knowledge I hope to collaborate with some of the semanticists from UMD’s strong linguistic group for which I have laid the foundation by taking relevant courses cross-listed in the linguistic department.

Semantics in Machine Translation

This idea of using a semantic formalism as an interlingua for translation has been explored previously as well. One of the early attempts was the UNITRAN system (Dorr ’87) that used a representation built on lexical conceptual structure as an interlingua. Approaches of this kind were the exceptions and most work in machine translation since the early 90s has been statistically driven. However in the recent few years there has been a growing interest in getting back semantics into the ball game, one of the main reasons being that the improvement using pure statistical based approaches seem to be reaching a plateau. The key idea behind incorporating semantics into statistical machine translation (SMT) is that it can enable the system to generate not only grammatical but also meaning-preserving translations. Lexical semantics can provide useful information for sense and semantic role disambiguation during translation. Compositional semantics can allow SMT to generate target phrase and sentence translations by means of semantic composition. Discourse semantics can help capture inter-sentence dependencies for document-level machine translation. Some examples: use of semantic role labels as a post process to reorder the output of traditional phrase based SMT (Wu et.al 2009), integrating semantic role features into SMT (Liu et.al 2010), use of lexical and semantic features for predicate translation and the reordering of arguments using semantic features (Xiong et.al 2012) etc. There have also been a few recent attempts of a more complete semantic MT system, e.g the use of synchronous hyperedge replacement grammar, a generalization of CFG from string to hyper graphs, to translate to and from graph structured meaning representation (Jones et.al 2012) and the use of lambda-calculus expressions as intermediate representation for translation (Andreas 2012).

Why AMR?

The key motivation behind developing AMR was to have a comprehensive and broad-coverage semantic formalism that puts together the best insights from a variety of semantic annotations (like named entities, co-reference, semantic relations, discourse connectives, temporal entities, etc.) in a way that would enable it to have the same kind of impact that syntactic treebanks (e.g. Penn treebank) have on natural language processing tasks. AMR tries to capture the notion of 'who did what to whom' in a sentence. It abstracts away from syntactic idiosyncrasies yet preserves the meaning of the sentence. It is composed of concepts and relations, not nouns and verbs. In fact the creators go as far as to say there are no nouns, verbs, adjectives, adverbs, affixes or zero pronouns in AMR. This allows AMR to have the same representation for a huge number of variations of a sentence (refer fig 1).

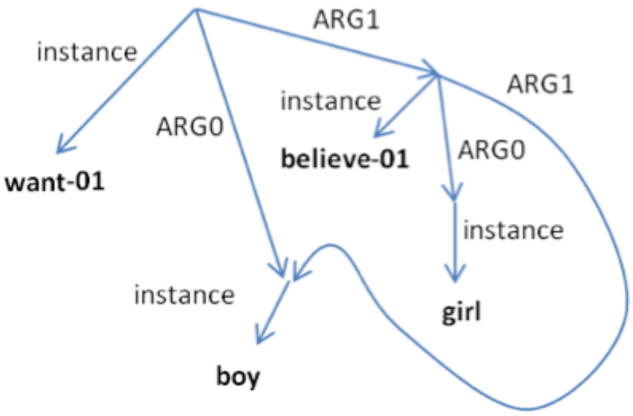

Figure 1: AMR above can be expressed variously in English as:
The boy desires the girl to believe him.
The boy desires to be believed by the girl.
The boy has a desire to be believed by the girl.
The boy's desire is for the girl to believe him.
The boy is desirous of the girl believing him.

AMR is built on top of PropBank (Kingsbury and Palmer, 2002; Palmer et al., 2005) and it uses PropBank's frames to define its concepts. PropBank has been found to be useful for many NLP tasks and it is being updated regularly. AMR itself has a steady growing annotated corpus, mainly for English along with a small annotated corpus for Chinese.

Use of AMR for multi-scale data

Most of the approaches currently require a huge amount of annotated data to train on and even after that it does not generalize well enough to work on unseen data. What I am proposing in this work is to train a system that uses a combination of annotated data and techniques built on linguistic intuitions that a human translator would use. This will not only help a system to generalize ~~more~~ across unseen data but also help translate better low resource languages. As noted in the NRT proposal - *"Working with 'small data' requires far more efficient approaches that generalize across domains by constructing more abstract language models".* AMR, by its very design, fits rightly under such an abstract model that would generalize well across domains and possibly across different languages.

References

Weaver, W. (1955). Translation. In *Machine translation of languages*, volume 14, pages 15–23. MIT Press, Cambridge, MA.

Bar-Hillel, Y. (1960). The present status of automatic translation of languages. In Alt, F. L., editor, *Advances in Computers*. Academic Press, New York.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1).

Dorr, Bonnie Jean. "UNITRAN: A principle-based approach to machine translation." (1987).

Wu, Dekai, and Pascale Fung. "Semantic roles for smt: a hybrid two-pass model." Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. Association for Computational Linguistics, 2009.

Liu, Ding, and Daniel Gildea. "Semantic role features for machine translation." Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010.

Xiong, Deyi, Min Zhang, and Haizhou Li. "Modeling the translation of predicate-argument structure for smt." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012.

Jones, Bevan, et al. "Semantics-Based Machine Translation with Hyperedge Replacement Grammars." COLING. 2012.

Andreas, Jacob. Toward Semantic Machine Translation. Diss. COLUMBIA UNIVERSITY, 2012.