Statement of Research Interests

My primary research interest is discourse in Statistical Machine Translation (SMT). My PhD topic is pronominal coreference in SMT – addressing the problem of translating pronouns within the framework of SMT, in which translation is performed at the sentence-level. Previous research on pronoun translation in SMT, including my own [1], has focused on third-person pronouns, treating them all as anaphoric (i.e. referring to an entity previously mentioned in the discourse). To date, little progress has been made on this problem and other types of pronouns have received very little attention. I believe that future progress is dependent on the careful analysis of pronouns as a whole. To facilitate this I developed the ParCor corpus as part of a collaboration with colleagues from the University of Edinburgh and the University of Uppsala. ParCor [2] is a parallel corpus of English and German texts in which pronouns are labelled according to their type (i.e. the function that they perform). Additional features are recorded for pronouns of different types. I have used the ParCor corpus to conduct analyses of manual translation, with the aim of identifying systematic differences in pronoun use between English and German, and therefore those pronouns for which future SMT research may be best directed. I have also assessed how well current state-of-the-art systems handle the translation of different pronoun types using a pronoun-specific manual evaluation task. Both analyses are presented in [3].

Given more time to work on the problem of pronominal coreference in SMT, there are several directions that I would wish to explore. These include further analyses of manual and automated translation and applying the knowledge that is gained to build pronoun-aware SMT systems. I believe that syntax-based SMT provides a logical starting point for handling intra-sentential anaphoric pronouns (i.e. those that appear in the same sentence as the entity they refer to), as these are syntactically governed. Another area which deserves attention is a deep analysis of the performance of the systems submitted to the DiscoMT 2015 shared task on pronoun translation, including my own automatic post-editing approach [4]. By identifying the strengths and weaknesses of each system, I hope to gain a better understanding of which models should be applied for specific pronouns. I would also be particularly interested in pursuing research on the topic of case prediction in German, with the initial view of applying it to the translation of pronouns.

I would also be very keen to work with other discourse-level phenomena, taking annotation, analysis and evaluation techniques developed for pronominal coreference and applying them to discourse connectives or lexical consistency, for example. Work on lexical consistency might begin as an extension of my earlier work on this topic [5].

I also have an interest in text simplification as a monolingual SMT problem. I was fortunate to be given the opportunity to explore this interest through proposing a Master’s project which eventually led to the development of two distinct projects for which I acted as co-supervisor under the guidance of Professor Bonnie Webber. These projects were hindered by the relatively small amount of parallel data available at the time, and this has not improved much since. Future work might therefore concentrate on the collation of more data.
References


