Introduction

It is a truism that texts have properties that go beyond those of their individual sentences, including:

- document-wide properties, such as topic mix, style, register, reading level and genre, all of which are manifest in the frequency and distribution of words, word senses, referential forms and syntactic structures;

- patterns of topical or functional sub-structure that show up in localized differences in the frequency and distribution of these elements within documents;

- patterns of discourse coherence, manifest in explicit and implicit relations between sentences (clauses), or between sentences (clauses) and referring forms, or between referring forms themselves;

- common use of reduced expressions that rely on context to convey a lot of information in very few words.

These properties stimulated a good deal of Machine Translation research in the 1990s, aimed at endowing machine–translated target texts with the same document and discourse properties as their source texts, albeit realized differently in source and target languages. This included work on stylistics for Machine Translation (DiMarco & Mah 1994), target language realization of source-language discourse relations (Mitkov 1993) and of referring forms (Bond & Ogura 1998; More et al. 1999; Wada 1990), anaphora resolution for generating appropriate target-language pronouns (Chan and T’sou 1999; Ferrández et al. 1999; Nakaiwa & Ikehara 1992; Nakaiwa 1999), and ellipsis resolution for generating appropriate target-language forms from ellipsed verb-phrases (Balkan 1998). Pointers to much of this work can be found in the Machine Translation Archive of conference and workshop papers from the 1990s (see www.mt-archive.info/srch/ling-90.htm).

This early period essentially ended with the 1999 publication of a special issue of the journal Machine Translation, edited by Ruslan Mitkov, devoted to anaphora resolution in Machine Translation and multilingual NLP. Only in the past 3–4 years has there been renewed interest in these topics, now from the perspectives of Statistical Machine Translation and Hybrid Machine Translation (Chung & Gildea 2010; Eidelman et al. 2012; Foster et al. 2012; Gong et al. 2011; Guillou 2012; Hardmeier & Federico 2010; Hardmeier et al. 2012; Le Nagard & Koehn 2010; Meyer 2012; Meyer et al. 2012; Voigt & Jurafsky 2012).

With this renewed interest, this ACL Workshop on Discourse in Machine Translation provides a timely forum for the presentation of new approaches to enabling modern systems to produce texts that are not merely sequences of isolated sentences.

Eight submissions have been accepted for the Workshop, on topics that range from multilingual modeling of discourse for machine translation, to actual use of discourse-level features to improve machine translation. From the modeling perspective, the papers presented at the Workshop discuss discourse phenomena such as lexical consistency (Guillou, this volume), lexical cohesion (Beigman Klebanov & Flor, this volume) and implicit connectives (Meyer & Webber, this volume), and “meaning units” with cognitive relevance (Williams et al., this volume). From the perspective of the application to MT, several papers present encouraging results showing that discourse-related features bring measurable improvements to the quality of machine-translated texts. One study uses oracle features, namely connective labels (Meyer & Poláková, this volume), while others use automatically-assigned ones. For instance, the translation of tensed verbs is improved by recognizing whether or not they are conveying
narrative material (Meyer et al., this volume); the translation of the pronoun “it” is improved based on lexical, syntactic and anaphoric features (Novák et al., this volume); and a document-level decoder is used when tuning an SMT system, with a sample of readability-related features (Stymne et al., this volume).

The studies presented at the Workshop provide quantitative data and benchmark scores to which future progress on these tasks should be compared. We hope that the Workshop will stimulate further work in these areas, as well as in the many areas of discourse and Machine Translation that are not yet represented.

We would like to thank all the authors who submitted papers to the Workshop, as well as all the members of the Program Committee who reviewed the submissions and delivered thoughtful, informative reviews.

Bonnie Webber (chair), Katja Markert, Andrei Popescu-Belis, Jörg Tiedemann (co-chairs)

References


Beata Beigman Klebanov and Michael Flor (2013). Associative Texture Is Lost In Translation. This volume.


Sara Stymne, Christian Hardmeier, Jörg Tiedemann and Joakim Nivre (2013). Feature Weight Optimization for Discourse-Level SMT. *This volume*.


Jennifer Williams, Rafael Banchs and Haizhou Li (2013). Meaning Unit Segmentation in English and Chinese: a New Approach to Discourse Phenomena. *This volume*. 