Semantic issues in Machine Translation

Abstract

Shortly after doing first steps towards finding ways for translating texts automatically, it became clear that ambiguity of words and structures is one of the hard problems, if not the one central hard problem of Machine Translation.

In this article we will describe the different shapes this problem takes (sections 1. Introduction, 3. Ambiguity). We will give a brief tour d'horizon about what types of architecture have been developed since the beginnings of MT and what kinds of solutions have been presented (section 2. History). We will concentrate on architectures which assign semantic representations to sentences and texts according to the Montegovian setting and which translate using such representations.

It has become common insight during the last decades that computing specific semantic representations is very costly and with respect to the needs in translation - where target sentences often may (and should) preserve the ambiguities of the source sentence - tends to be a kind of overkill. Therefore, shallow semantic representations will often do, but not always, as we will try to motivate.

Because of its turning to referential elements in texts, DRT is especially suited for the representation of texts and, therefore, for being used in semantically based Machine translation, especially with respect to translating referential elements like pronouns and tenses.

There have been developed a number of semantic formalisms since the early nineties which allow for compact representations of semantic ambiguities by so-called underspecified structures (section 4. Semantic representation and underspecification). We want to introduce relevant representatives of these formalisms briefly and, then, concentrate on an underspecified descendant of DRT, so called flat underspecified DRT (FUDRT) which develops from UDRT (underspecified DRT) by incorporating a number of additional features for the representation and dynamic evaluation of different types of ambiguity, like lexical and functional ambiguity. We will illustrate how such a formalism can be integrated in a typical transfer architecture, how it contributes to optimise modularisation of the system by allowing to represent most types of idiosyncratic word-specific transfer instructions in the bilingual lexicon (keeping the general transfer procedure free from them), how it contributes to defining such instructions - which are: conditions of appliance and transfer- (i.e. restructuring-)statements and how such conditions may trigger dynamic evaluation of the content of partial representation structures, thus deciding about the correct choice of alternative target words or structures. Using a number of examples from different areas of cross-linguistic mismatches (lexical, ambiguity, referential ambiguity, scopal ambiguity, attachment ambiguity), we try to motivate that resolving ambiguities is an interleaved procedure, where a specific resolution of some ambiguity may constrain the possible resolutions of another one (sections 5. Underspecified representation, 6. Lexicalist, recursive transfer of underspecified representations).

Though practical, this type of transfer architecture is still costly, because it needs a lot of knowledge to be encoded in the system.

Recent approaches to Machine Translation try to do without any predefined knowledge by learning translation relations from huge bilingual corpora using statistics based MT systems being ignorant at the beginning. Often it is argued that this type of translation allows for

economic and fast definition of robust MT systems. In any case, a disadvantage is that, beyond a certain limit, it is difficult to tune such systems for better quality. Therefore, it seems promising to think about integrated approaches which take over advantageous features of the competing architecture. With respect to rule-based systems, it seems advantageous to incorporate statistics based methods for (semi-)automatic learning of new (bi- and monolingual) lexical relations from corpora and for determining suitable weights measuring the relevance of relations and structures. Given this perspective of adding statistically gained information to semantics-based transfer systems, underspecified representations seem to be especially practical, since they are more analytic than conventional representations, subdividing the one representation into parts and thus providing (more) relations between the parts of the sentence information (and functions defined on them) which, in a natural way, can serve as interfaces for receiving information about weights and evaluation preferences (section 7. Flat, underspecified rule-based MT and current empirical MT-trends). The paper will end with a short concluding summary and outlook (section 8. Conclusion).

1. Introduction
    - 1.1 Purpose of semantic information in Machine Translation
    - 1.2 Ambiguity-preserving translation

2. History
    - 2.1 A brief history of the common Machine Translation Architectures
    - 2.2 The role of semantic representation in different traditional MT architecture types
    - 2.3 Trends in current MT research and development.

3. Ambiguity
    - 3.1 Types and complexity
    - 3.2 Processing constraints in MT

4. Semantic representation and underspecification
    - 4.1 Semantic representation theories
    - 4.2 Underspecification formalisms

5. Underspecified representations
    - 5.1 Representation of different types of ambiguity
    - 5.2 Analysis modules
    - 5.2 Dynamic semantic evaluation
    - 5,3 Resolution triggers

6. Lexicalist, recursive transfer of underspecified representations
    - 6.1 Transfer module
    - 6.2 Features of a lexicon formalism
    - 6.3 Dynamic evaluation

7. Flat, underspecified rule-based MT and current empirical MT-Trends
    - 7.1 Learning from corpora
    - 7.2 Learning selectional restrictions
    - 7.3 Learning structural preferences
    - 7.4 Learning contextual restrictions of transfer equivalents

8. Conclusion