The Gain of Failures: Using Side-Effects of Anaphora Resolution for Term Consistency Checks

Galia Angelova¹, Ognyan Kalaydjiev¹, and Walther von Hahn²

¹ Bulgarian Academy of Sciences, 25A Acad. G. Bonchev Str.,
1113 Sofia, Bulgaria
{galja,ogi}@iml.acad.bg
² University of Hamburg, Vogt-Koelln-Str. 30
D-22527 Hamburg, Germany
vhahn@nats.informatik.uni-hamburg.de

Abstract. The paper discusses a prototype module for on-line checking of term consistency in a workbench for knowledge-based Machine Aided Human Translation (MAHT). We present the linguistic resources and the knowledge base (KB) of the system as well as their place in the processes. To discover missing or misleading translations, the checker relies on the lexicon information and the hierarchy in the KB. To detect comprehension difficulties, the module checks ambiguities in the target text by resolving anaphora. The module design is based on the assumption that in aligned parallel texts some types of difficulties to resolve anaphora could be used to find wrong, missing and/or cognitively difficult translations. The paper explains the algorithms for checking term consistency. An evaluation of the approach and prospects for further work conclude the paper.

1 Introduction

The development of user friendly workbenches for Machine-Aided Translation (MAT) requires the integration of flexible linguistic algorithms for a more intelligent system behaviour. Additionally, such algorithms should be hidden from the user translator in MAHT, since the typical human user is not acquainted with computing details and therefore (s)he should always see a well-known software environment, similar to e.g. the every-day word processors (s)he is accustomed to.

The review [3] classifies three types of present MAHT systems in real use:

(i) for professional translators working in a team: such MAHT systems provide access to bilingual terminology, to a translation memory and submit selected phrases to Machine-Translation (MT) systems;

(ii) for independent professional translators: these systems combine large multilingual lexicons of different types with a translation memory;
for occasional translators: in addition to their lexicons, these systems support target templates (with ready-made text fragments) for letters and other kinds of standard documents in the target language.

None of these systems includes advanced NLP techniques, except MT background systems in case (i), but this software is only available in few large translation organisations.

This paper discusses how anaphora resolution can be exploited for checking the correct translation of terminology in a MAHT system. The checker (called TETRARCH) is being developed in the experimental MAHT environment of DBR-MAT\(^1\). The general design of this module is based on the following idea: After an (automatic) alignment of two parallel texts, term consistency can be verified firstly by checking lexical consistency among the terms and their translations and secondly by checking anaphoric consistency by comparing sequences of full terms, generalised lexical abbreviations/replacements and pronominalisations in both texts. Thus accuracy, consistency and readability of translations can be increased.

Checking such phenomena while the user is writing the (draft of the) target text would trigger on-line warnings about the quality of the translation. Thus we integrate another intelligent component into a pilot MAHT-system. It does not perform anaphora resolution but is rather drawing the attention of the user to inconsistencies by exploiting ambiguities in anaphora resolution.

The paper is structured as follows. In Sect. 2 we discuss the general DBR-MAT paradigm, i.e. its background of linguistic and knowledge resources in which the TErm TRAnslation CHecker (TETRARCH) is located. Sect. 3 considers briefly techniques for resolution of pronominal anaphora which can be applied in our approach. Sect. 4 describes TETRARCH using German-Bulgarian and German-English examples. Sect. 5 evaluates the application of TETRARCH in the DBR-MAT system, which supports translations from German to Bulgarian. Sect. 6 summarises the paper and presents some conclusions.

2 The DBR-MAT Environment

The DB-MAT and DBR-MAT projects investigate a new MAHT paradigm where the human user is supported by linguistic as well as subject information [14], [16]. Domain knowledge is not encoded in the lexicon entries but in a single, language-independent KB of Conceptual Graphs (CG), from where explanations are generated in different natural languages (NL) [1]. The translator triggers the generation of explanations by highlighting a text fragment and choosing a query from a nested menu, similar to usual menus in text-processing systems. The user can receive recursive explanations about any knowledge in the KB; thus DB-MAT represents a NL interface for browsing domain knowledge [2]. The DB-MAT demo system is designed for technical documents and illustrates an

\(^1\) DB-MAT (1992-95) and DBR-MAT (1996-98) projects investigate a knowledge-based MAT paradigm. Funded by the Volkswagen Foundation (Germany), see http://nats-www.informatik.uni-hamburg.de/dbrmat/db-mat.html