Homework 4: Project-Related Paper Report

Christopher Brown
chrisbrown@utexas.edu

April 17, 2013

1 Summary

Siegel (1996)[1] addresses the specific problem of Japanese to German machine translation, focusing on Japanese features that are not manifested at surface level. Japanese morphology does not consistently show plural status, nor does Japanese exhibit number agreement between nouns and verbs. Japanese plural markers are optional and can only be applied to noun phrases referring to people. German determiners like *ein* and *dem* do not align to anything in the source Japanese sentences, but they are required in the German translation. Siegel's goal is derive rules for inserting appropriate determiners in the German sentence output.

It's not a simple matter of post-processing output in the target language; Siegel claims several factors affect the choice of determiners in the German:

- Japanese surface level
- German lexical restrictions
- Domain / discourse restrictions

Together, these determine which insertion is most appropriate. Siegel asserts that the general problem is 'interlingual,' and must be addressed at the level of translation.

Other approaches have included pre-processing Japanese text for relevant markers, inferring definiteness and number at the source language level. Siegel claims this is too shallow. For example, *shain* (staff member) can translate to either *die Mitarbeiter* (the staff-members) or *der Mitarbeiter* (the staff-member), and this is 'dependent on the domain.' She discusses ways to translate the number and definiteness using Prolog transfer rules separate from the content, so that rules regarding number / definiteness are more general than the lexical transfers seen in just the observed data.

While Siegel denies that inferring definiteness is inherently either a Japanese problem or a German problem, she claims that inferring gender at the German level is inherently a German problem, despite systematic / typological gender patterns in some languages that exhibit gender on nouns.

Some Japanese nominal phrases contain a determiner (*sono*/*kono*) that translates to *these*, and predictably marks definiteness. Japanese numerals behave normally and are clear indicators of number, when they occur. Siegel proposes a rule that transforms a determiner-led noun phrase without number into a definite German noun phrase, and suggests that similar rules could use other Japanese features, specifically certain adjectives and genitives.

Siegel's corpus is small; 10 dialogs, which includes a total of 566 noun tokens at the Japanese level. Her transfer rules seem to be derived from observations solely on this corpus—not on any greater familiarity with translating Japanese to German. Using a combination of preference rules with defaults, Siegel shows how an assortment of transfer can handle translating several types of Japanese noun phrases into German, inserting number and definiteness markers that are not given in the Japanese source sentences.

[1] It is an ancient paper, in NLP-years, but there is very little current research on determiners. The methodology may be obsolete, but the issues are still relevant.

2 Improvements

Successfully implementing transfer rules for translation requires deep understanding of language, which isn't feasible given current databases and available linguistic infrastructure. Simpler models that use more data are more successful, as most modern machine translation systems demonstrate.

The semantic issues arising in Siegel (1996) are interlingual problems, as the author claims. One could argue that all aspects of translation are interlingual, but machine translation techniques ignoring the theoretical complexities often surpass more linguistically correct systems. All models are wrong, but some are useful; approaches like Siegel's are more 'correct,' as far as Prolog-driven models go, but they aren't as useful.

Siegel's methodology is highly domain specific; not only are the rules she proposes specific to translating from Japanese to German, but even these are tuned for the translated informal conversation corpus that she discusses. Many more languages than Japanese lack determiners; many more than German have them. It would be preferable to start with research considering the problem at a more general level. The rules she proposes may have correlates in other language pairs; I don't know how transferable these transfer rules are, but their lexical specificity suggests that even if the rules would be similar for, say, English, they are not general enough to extended across languages automatically.

Evaluation

Siegel does not demonstrate that her transfer rules are better at translating than simpler pre- or post-processing rules. This research is very early in machine translation, before the statistical revolution, but she relies on intuition rather than results.

My project investigates how much can be achieved via non-aligned surface-level models. It is not only a translation problem, but one aspect of semantics which has a major application in translation, particularly by improving post-processing of the translation system's output. Comparing this to a baseline, my approach could itself become a baseline for claims like Siegel's, since I address just one of the three factors that Siegel claims this problem depends on. If her approach does significantly better than the post-processing of my model, that would be substantial support for her claim of interlingual dependence, and justify further research into inferring definiteness placeholders on the source language, so that they can be translated.

Conclusion

Siegel doesn't discuss translating German back to Japanese; is it sufficient to simply drop most German determiners and number morphology? If so, are there easy ways to predict which things get dropped? Are there other issues that arise in the interaction between these two systems? Further, I'm curious if Japanese–Russian translation is easier, because those languages both lack determiners.

My project is a cross-section of some of issues raised in Siegel (1996). Instead of looking at a particular language pair, I'm investigating the recoverability of definiteness markers at the surface level—on the target language, in a translation scenario. This is more general; whereas language pairs are quadratic in the number of languages handled by the translation system, and translation between them requires a large number of parallel texts, surface-level post-processing depends only on the target language and does not require the same depth of annotation. This means much more data can be put to use, more easily.

References

Melanie Siegel. Preferences and defaults for definiteness and number in Japanese to German machine translation. In *Language, Information and Computation (PACLIC 11)*, pages 43–52, 1996.