Substring Alignment using Suffix Trees

Martin Kay

Stanford University

Abstract. Alignment of the sentences of an original text and a translation is considerably better understood than alignment of smaller units such as words and phrases. This paper makes some preliminary proposals for solving the problem of aligning substrings that should be treated as basic translation unites even though they may not begin and end at word boundaries. The proposals make crucial use of suffix trees as a way of identifying repeated substrings of the texts that occur significantly often.

It is fitting that one should take advantage of the few occasions on which one is invited to address an important and prestigious professional meeting like this to depart from the standard practice of reporting results and instead to seek adherents to a new enterprise, even if it is one the details of which one can only partially discern. In this case, I intend to take that opportunity to propose a new direction for a line of work that I first became involved in in the early 1990's [3]. It had to do with the automatic alignment of the sentences of a text with those in its translation into another language. The problem is non-trivial because translators frequently translate one sentence with two, or two with one. Sometimes the departure from the expected one-to-one alignment is even greater. We undertook this work not so much because we thought it was of great importance but because it seemed to us rather audacious to attempt to establish these alignments on the basis of no *a priori* assumptions about the languages involved or about correspondences between pairs of words.

As often happens, it turned out that what we thought of as new and audacious was already "in the air" and, while we a were at work, Gale and Church[2] published a solution to the problem that was arguably somewhat less accurate than ours, but was altogether simpler, computationally less complex, and entirely adequate for practical purposes. Whereas their approach was based on nothing more than the lengths of the sentences, in terms either of characters or words, ours made hypotheses about word alignments on the basis of which it circumscribed the space of possible sentence alignments. It then refined the initial set of word alignments, and proceeded back and forth in this manner until no further refinements were possible. Fortunately the process converged fast.

In the relatively short time since this work was done, sentence alignment has come to be seen as a central tool in work on machine translation. During this time the perception has grown that the rules that direct the operation of a machine-translation system should be derived automatically from existing translations rather than being composed one by one by linguists and system designers. The first step in just about any such learn-