Language Weaver: The Next Generation of Machine Translation

Bryce Benjamin, Laurie Gerber, Kevin Knight, and Daniel Marcu

Language Weaver, Inc.
4640 Admiralty Way, Suite 423
Santa Monica, CA 90292
{cbryceb, lgerber, knight, marcu}@languageweaver.com

Abstract

We introduce a new generation of commercial translation software, based primarily on statistical learning and statistical language models.

1 System category: Commercial Product

Language Weaver will begin licensing server-based machine translation software in September 2003. In addition to software licensing, Language Weaver offers system customization services.

2 Hardware platform and operating system

Windows servers and Windows XP standalone systems.

3 Language Weaver Software

Language Weaver is reducing the cost of high-volume, domain-specific translation. Two primary factors have limited the deployment of MT for this type of translation in the past: the cost of customizing system dictionaries with the required terminology, and the unnatural, inflexible style and poor quality of the translated output. Language Weaver's proprietary statistical learning technology addresses these points with:

Automated customization: The system learns to translate new language pairs and new subject areas and text types automatically. The learning process captures the terminology and style of the subject area and text type from the translations used as training material. While Language Weaver can learn from existing translations, such as those in translation memories, Language Weaver is not a translation memory system. Once it has completed its learning process, it is a fully automatic machine translation system, and can translate previously unseen text.

High quality translations: Language Weaver offers the opportunity to generate high quality translations following customization. The more learning material available, the higher the resulting translation quality will be.

Natural, appropriate output style: Because the learning process captures both terminology and style, Language Weaver's target language output is extremely natural-sounding, and appropriate to the text. This makes it possible to cost-effectively postedit Language Weaver translations if polished translations are required.

Handles non-standard language: Many applications involve controlled language or text written in a style that is unique to the application. Even product documentation tends to involve idiosyncratic grammar and usage, as it strives to succinctly describe the product to users. Because Language Weaver learns from the client's texts, it is agnostic about grammar and usage. The core algorithms are language-independent. It learns whatever it is exposed to.

4 Custom Translation Systems

The first product planned for release by Language Weaver in 2003 will be a fully automatic client-server machine translation system for Arabic, Chinese, French, Hindi, Italian and Spanish <> English, accessible through a flexible API. Language Weaver customizes each translation

system to the customer's domain(s). The customization of a language pair system for a new customer domain, involves the following steps:

Parallel corpus utilization: In the first step, a data set (corpus) of parallel texts is utilized. This corpus consists of source and target language texts covering the same subject area and type of text that the resulting translation system will be used for. (A translation memory database would be an example of an existing corpus.)

Automatic training: Once a large set (1 million words or more) of source and target text segments is prepared, only a short time is required to derive the statistically weighted translation patterns (parameters) from the training data.

Decoder/Translator: Once the statistical parameters have been learned, they can be used by the *decoder* to translate new text that was not in the training corpus.

5 Conclusion

Language Weaver offers a new approach to machine translation that is a significant advance in automated translation.

6 About Language Weaver

Language Weaver was incorporated in January, 2002 to commercialize a statistical approach to automatic translation. Language Weaver's proprietary statistical translation technology is the result of twenty person-years of invention and development at the University of Southern California's Information Sciences Institute (USC/ISI) by Dr. Kevin Knight and Dr. Daniel Marcu. Ongoing research funding for their work at USC/ISI is provided by several government agencies including the Defense Advanced Research Projects Agency (DARPA) and the National Science Foundation (NSF).

7 Bibliographical References

Daniel Marcu and William Wong (2002). "A Phrase-Based, Joint Probability Model for Statistical Machine Translation." Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002), Philadelphia, PA, July 6-7

Marcu, D. 2001. "Towards a Unified Approach to Memory- and Statistical-Based Machine Translation." Proceedings of ACL-01. Toulouse, France.