Cairo: An Alignment Visualization Tool

Noah A. Smith* **and Michael E. Jahr†**

*University of Maryland
College Park, Maryland, USA
nasmith@cs.umd.edu

†Stanford University
Stanford, California, USA
mjahr@stanford.edu

Abstract

While developing a suite of tools for statistical machine translation research, we recognized the need for a visualization tool that would allow researchers to examine and evaluate specific word correspondences generated by a translation system. We developed Cairo to fill this need. Cairo is a free, open-source, portable, user-friendly, GUI-driven program written in Java that provides a visual representation of word correspondences between bilingual pairs of sentences, as well as relevant translation model parameters. This program can be easily adapted for visualization of correspondences in bi-texts based on probability distributions.

1. Background

Cairo was originally developed with statistical translation models of the sort developed at IBM in the early 1990s in mind (Berger et al., 1994). These models, referred to as the "Candide" models, are based on the source-channel paradigm, where one language (the "source" language) is interpreted as an encoded form of another (the "target" language)[1] (Weaver, 1949). A translation model is built using iterative training on a sentence-aligned bi-text (Brown et al., 1993). Translations are produced through a process known as "decoding" (Wang and Waibel, 1997).

Statistical translation systems are difficult to understand, even for their designers, because of the vast number of parameters learned automatically. Candide Model 3, for example, consists of four types of parameters to model uni-directional language translation. Using the Italian-to-English translation example, these are:

Translation: The probability that an English vocabulary item (type) will translate to a given Italian type.

Fertility: The probability that a given English type will translate into n Italian tokens.

Distortion The probability that any English token in a given indexed position i of an English sentence will align with an Italian token in given position j of an Italian sentence, given the lengths of the sentences.

NULL-insertion The probability of insertion of a NULL token at any position in the English sentence during the encoding (a NULL token is empty in English but is aligned to one or more Italian tokens). This is a scalar parameter for the model.

[1]The source-target terminology can be counter-intuitive when discussing source-channel models. If we are building a machine translation system to translate sentences from Italian to English, then let Italian be the "source" language and English the "target" language. To avoid confusion, we will use the Italian-to-English example exclusively in this paper.

Cairo was implemented to allow inspection of the bilingual text word alignment process and the decoding process in statistical machine translation models.

2. Capabilities

Cairo takes as its input an SGML-style file that specifies the two sentences, their alignment[2] (in a simple format), and all relevant model parameters. Relevant parameters may include the probabilities associated with a word translation, fertility, etc. that actually occur in the alignment as well as other probabilities to be used as a basis for comparison. For example, in the alignment in Figure 1, the English token *red* has a fertility of 1, because it corresponds with one Italian token, *rosse*. The fertility table on the right displays the probability that *red* would have this fertility, and it also displays the probability that the fertility would be 0 or 2. The more information given as input, the more powerful the visualization will be; however, no information is required except for the sentences themselves.

In a graphical user interface (GUI), Cairo displays the given sentence pair (assumed to be a translation pair) with lines drawn between aligned words. This representation can be displayed vertically or horizontally, and accommodates sentences of arbitrary length, using scrolling windows. (See Figure 1.)

Each token in a sentence can optionally include multiple, parallel streams of data. One common use by our collaborators is to include part-of-speech tags, lemmas, and/or most-probable translations for each token. (See Figure 2.) The number and names of all streams are provided in the input file by the user. The user has full control over which streams are displayed.

Cairo allows for up to three alignments to be displayed at once for a sentence pair. Visual rendering of multiple alignments is accomplished with different colors, using subtractive color mixing to indicate shared word correspondences.

[2]We will use the term "alignment" to refer to the set of all word correspondences between a sentence pair.

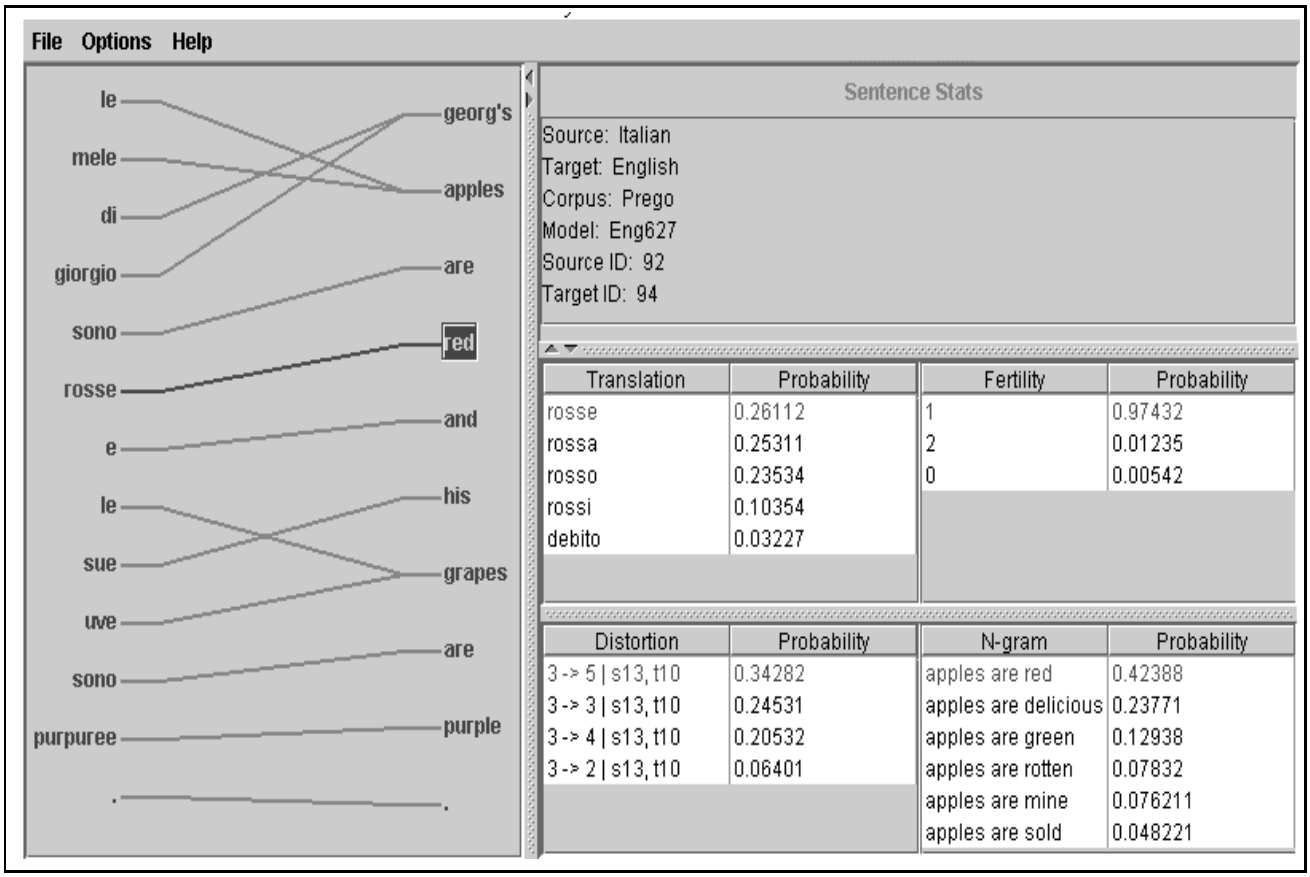

Figure 1: Cairo displaying an aligned sentence pair vertically. The word *red* is selected, and some of its respective translation and language model probabilities are shown in the tables to the right. The probabilities relevant to the translation shown are actually displayed in red, as are the corresponding Italian word *rosse* and the linking line.

When evaluating a machine translation, it is useful to refer to a gold standard, or reference translation and alignment. Cairo provides the simultaneous display of a reference translation alongside the given translation, and allows rapid switching of the display between the machine alignment and the reference alignment.

A Cairo user can mouse-click on an English token and see its model parameters displayed in tables alongside the alignment. In addition, if relevant language model parameters (used in tandem with the translation model in the decoding process) are specified, then these will be displayed as well. For example, if the user clicks on an English word ε, a list of Italian words to which ε is likely to translate appear in the tranlsation table, sorted by probability. The Italian words $\iota_1, \iota_2, \ldots, \iota_n$ which are aligned to ε for each alignment are displayed in colors matching their respective alignments. Likewise, sorted lists of fertilities, distortions, and likely English words (*i.e.*, expected by the language model) are displayed. One, two, or four of the tables may be shown at a time.

A window displays information such as the names of the languages the sentences are in, sentence identification numbers, the language model used, etc. Any such text given in the input file is displayed in this window.

3. Use

Cairo has proven extremely useful in statistical machine translation research. When dealing with models consisting of millions of parameters, actually examining specific alignments can be enlightening. Being able to see potential neighborhood alignments by looking at other model parameters that "might have been" gives additional insight. This tool has allowed us and others to track the progress of translation models and discover models' biases that affect performance.

During experimentation with the Egypt implementation of the Candide machine translation system, Cairo was used to view correspondences in bilingual sentence pairs (Egypt generates Viterbi word alignments for both the training and test corpora) (Al-Onaizan et al., 1999). One of the discoveries made, using Cairo, was a tendency of rare word types to have high fertilities. That is, this class of translation models will cause many source-language (e.g., Italian) tokens (usually high-frequency and closed-class) to correspond to a single low-frequency target (e.g., English) word. This is an important effect that would have gone unnoticed if our exploratory tools had been limited to tables of learned statistical parameters and plain-text depictions of word correspondences between sentence pairs.

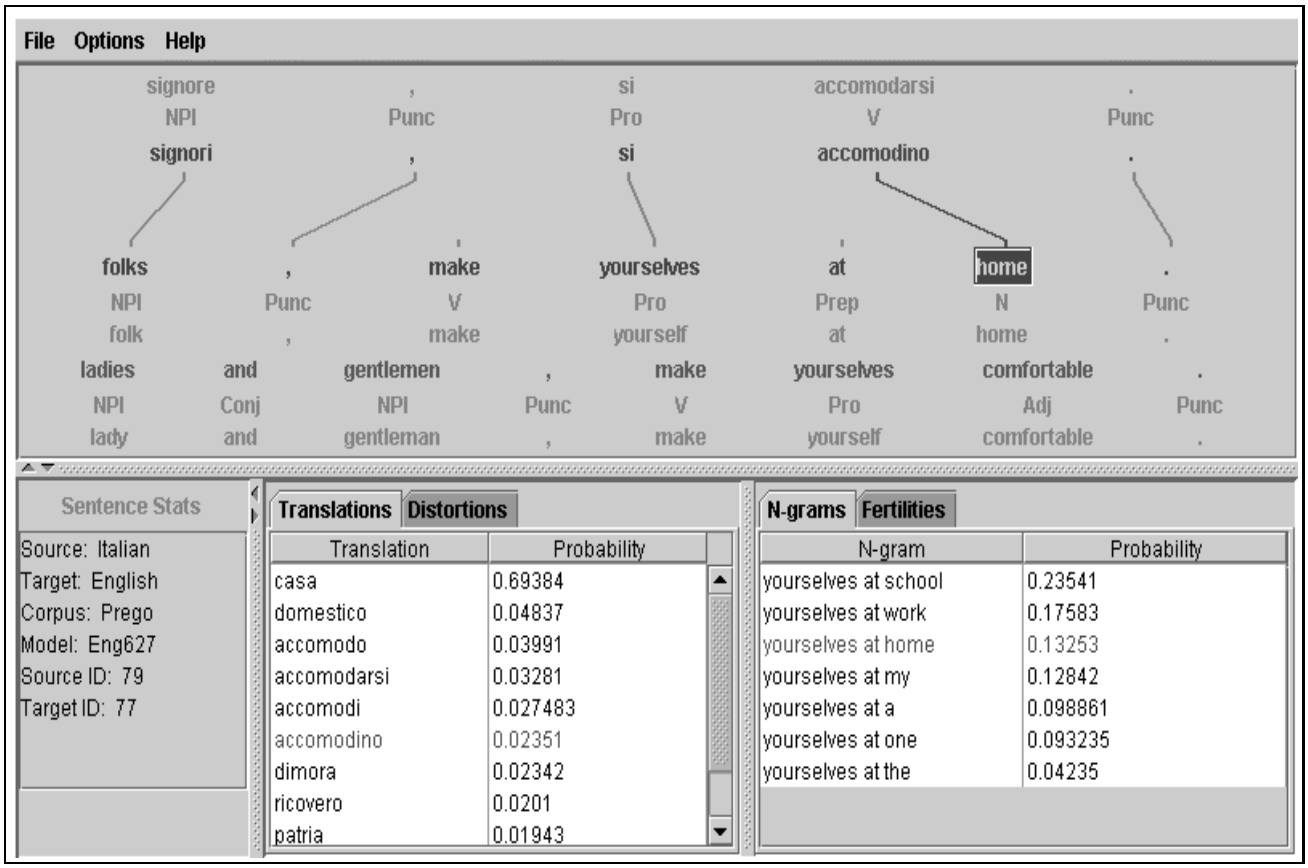

Figure 2: Cairo displaying an aligned sentence pair horizontally. The word *home* is selected, and some of its relevant translation and language model probabilities are shown in the two tables in the bottom right. A reference translation is shown below the machine translation. Note also the multiple streams, used here to show parts of speech and lemmas.

4. Implementation

Cairo is implemented in Java, using the Swing interface library (Sun Microsystems, 1999); this makes it portable to any architecture and useful in Web-based applications. Cairo is freely available at http://www.clsp.jhu.edu/ws99/-projects/mt/toolkit/cairo.tar.gz, alongside the Egypt statistical machine translation toolkit (Al-Onaizan et al., 1999). Also available is a script which converts the output of the Egypt translation model training tool into Cairo's SGML format. It is possible to modify and extend Cairo for purposes other than statistical machine translation.

5. Acknowledgements

Cairo was developed at the Johns Hopkins University Center for Speech and Language Processing 1999 Workshop in Language Engineering. We would like to express gratitude to the Statistical Machine Translation team at the Workshop for helping to define this tool, and in particular Kevin Knight and Dan Melamed. In addition, Philip Resnik has provided valuable advice in this effort.

6. References

Al-Onaizan, Yaser, Jan Cuřin, Michael Jahr, Kevin Knight, John Lafferty, I. Dan Melamed, Franz Josef Och, David Purdy, Noah A. Smith, and David Yarowsky, 1999. Statistical machine translation. Technical report, Johns Hopkins University Center for Language and Speech Processing.

Berger, Adam L., Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John R. Gillett, John D. Lafferty, Robert L. Mercer, Harry Printz, and Luboš Ureš, 1994. The candide system for machine tranlsation. *Human Language Technology*:157–162.

Brown, P. F., V. J. Della Pietra, S. A. Della Pietra, and R. L. Mercer, 1993. The mathematics of statistical machine translation: parameter estimation. em Computational Linguistics 19(2).

Sun Microsystems, 1999. Creating a GUI with JFC/Swing. http://www.java.sun.com/docs/books/tutorial/uiswing/.

Wang, Ye-Yi and Alex Waibel, 1997. Decoding algorithm in statistical machine translation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*.

Weaver, Warren, 1949. *Machine Translation of Languages*. Massachusetts Institute of Technology Press.