Combining Resources for Open Source MT

Eric Nichols,♯ Francis Bond,♮ Darren Scott Appling,♭ Yuji Matsumoto♯

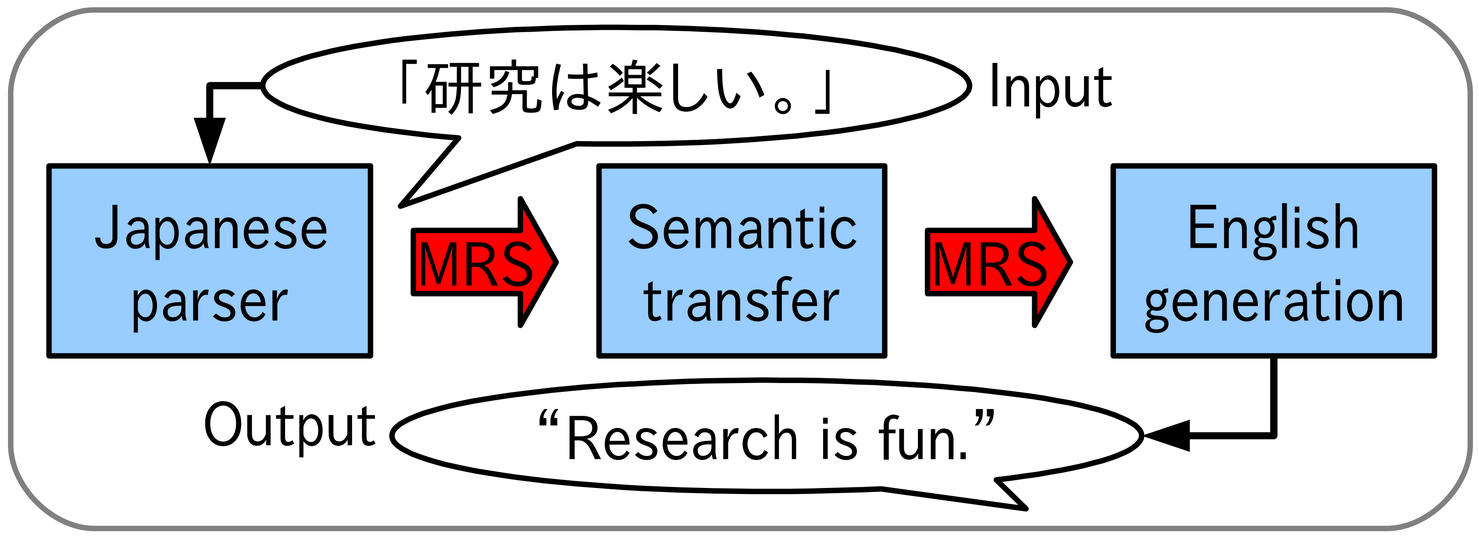
Overview

Many machine translation systems have been developed, but few are available as open resources. This makes it difficult for researchers to share resources, learn from other systems, and pool efforts. We present a hybrid system that combines rule based semantic-transfer and statistical translation engines that is made entirely of open source components. The HPSG grammars and rewrite engine used in semantic transfer are from the DELPH-IN project, and the SMT component uses the Moses system. Rules for the semantic transfer engine are mined from the Japanese→English data in JMDict, a freely available multilingual dictionary.

Other Open Source MT Systems

Open source MT systems are important because they make it easier for researchers to collaborate and learn from each other's approaches. Recently, a few open source MT systems have been released.

OpenTrad is a shallow transfer system that focuses on Spanish, but has been expanded to handle many language pairs. OpenLogos is a rule-based system of commercial origin that has recently been opened. In the SMT community, Pharaoh, Moses, and SRILM represent efforts to bring statistical tools to the user.



Jaen: Semantic Transfer Based MT

We have developed a Japanese→English semantic transfer based MT system using the LOGON engine. Jaen uses a common semantic formalism, Minimal Recursion Semantics (MRS), for both analysis and generation. An HPSG parser analyzes the source language sentence producing an MRS structure. Translation rules rewrite the source MRS into a target language MRS structure that is used for generation. The system uses a hierarchy of hand-crafted rule types to simplify development.

Dictionary-based Rule Acquisition

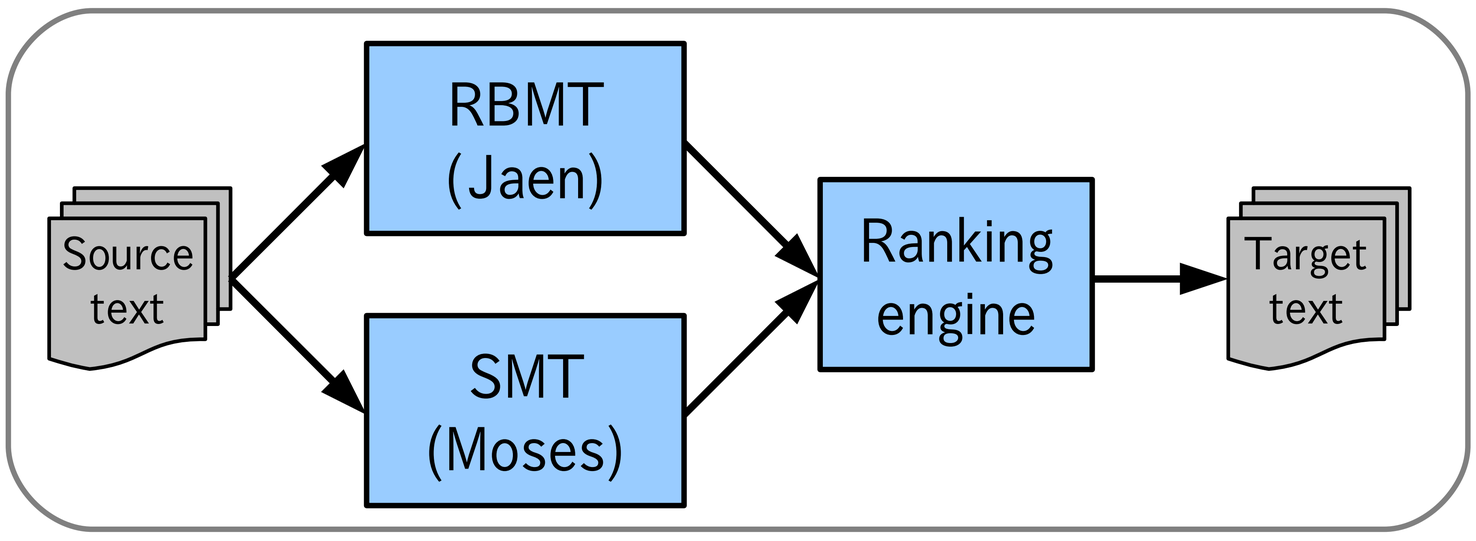
We automatically acquire transfer rules from translations pairs obtained from various sources. We filter translations against a collection of hand-crafted mappings identifying the rule types compatible with the syntactic categories of a translation's words. Then we format valid entries into transfer rules. We currently have acquired over 50,000 rules.

Corpus-based Rule Development

In order to bootstrap the development of Jaen-specific transfer rules, we parsed sentences from the BTEC* corpus and attempted to translate them using only the rules we acquired from JMDict. We identified the semantic relations that did not transfer and wrote rules for the most common. Currently, we have close to 200 hand-crafted rules.

Example Translations

Jaen	Moses	Reference
Are Japanese dogs big?	It is a big dog in Japan?	Are Japanese dogs big?
Where is there a Nihon embassy?	Where is the Japanese Embassy?	Where is the Japanese Embassy?
Is there a hotel in this vicinity?	Is there a hotel near here?	Is there a hotel around here?
A center hotel.	The hotel.	The Center Hotel.
Did you see criminals?	Did you see the?	Did you see who did it?
Abdomens hurt.	腹部 aches.	I have a stomach ache.
Please do an allergy check.	I am allergic to check, please.	I'd like to have an allergy test, please.
Is it a front money government?	Do I need to pay in advance?	Do I need to pay in advance?



Open Source Resources

- DELPH-IN <http://www.delph-in.net>
  - Jacy, a Japanese Grammar
  - English Resource Grammar
  - LOGON Semantic Transfer Engine
- JMDict: a Japanese-Multilingual Dictionary
- Moses <http://www.statmt.org/moses>

Contact

♯ {eric-n,matsu}@is.naist.jp
Nara Institute of Science and Technology
Computational Linguistics Laboratory
♮ bond@ieee.org
National Institute of Information and Communications Technology
♭ darren.scott.appling@gatech.edu
Georgia Institute of Technology

Discussion

Currently there are few sentences in our evaluation data where Jaen and Moses can be directly compared, so we do not give a quantitative analysis at this point. However, a qualitative analysis is still useful. In the above examples, the system with the best output is colored. As may be expected, Jaen does a better job of preserving the syntactic structure of a translation, where Moses is better at learning translations in context.

Moses: Open Source Statistical MT

Moses is an open source statistical machine translation system that makes it easy for users to make their own systems. It offers a beam-search decoder and tools for learning alignments. We constructed a simple Japanese→English SMT system using the data set from the 2006 International Workshop on Spoken Language Translation. It is incorporated into our system using a ranking engine, which is currently a simple cascaded model: Jaen results are returned whenever present, falling back to Moses otherwise.