Language Technologies in e-Commerce

Lecture #7
Machine Translation (1)

http://www.cs.cmu.edu/~ralf/lectures.html

Classification of MT Systems

By depth of analysis:

- direct transfer
- transfer
- interlingual

By knowledge source:

- Rule-Based ("Traditional")
- Data-Driven
  - statistical
  - example-based
  - neural networks

By Parsing Algorithm:

- Chart Parser
- Augmented Transition Networks
- Context-Free Grammar
- Lexical Functional Grammar
- Phrase-Structure Grammar
- Conceptual Dependency
- Semantic Grammars
- Slot Grammars
- ...

Early MT

Alan Turing, 1948:
Computers could be used for

> "(i) Various games, e.g. chess, noughts and crosses, bridge, poker; (ii) the learning of languages; (iii) *translation of languages*; (iv) cryptography; (v) mathematics."

Warren Weaver, 1949:

> "When I look at an article in Russian, I say, 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'"

Also speculated on the existence of prelinguistic symbols underlying all human thought; if languages could be converted into such a code, it could become a universal interface between languages.

First US MT Conference, 1952

Georgetown University, January 1954:
First public demonstration of MT (Russian to English)

- 250-word vocabulary
- 6 grammar rules
- 49 hand-picked sentences

First International MT Conference, 1956

Early MT (2)

The earliest systems were direct-transfer systems: replace words and phrases by their translations and apply a bit of re-ordering to the output.

> *Ah, well, I don't want you to get the impression it's just a question of the number of words... um... I mean, getting them in the right order is just as important.*
>
> -- John Cleese as Sir Edwin, the great actor

Advantages of Transfer

- no need to define language-independent representations
- avoid complexity of deep analysis

Disadvantage of Transfer

- need a complete new system for each language pair

TAUM-METEO

Restricting the domain of discourse to a narrow subject makes translation much easier.

Weather reports and forecasts are a very narrow domain.

The Canadian METEO system has been translating weather bulletins between English and French since 1977.

- 220-word (2000 phrases?) vocabulary
- restricted syntax
- 98% correct translations

(TAUM = Traduction Automatique de l'Universite de Montreal)

METAL

"Machine Translation and Analysis of Natural Language"

Commercial multilingual machine translation system for volume translations of technical material; by Siemens-Nixdorf and Linguistic Research Center (University of Texas, Austin)

modified syntax-oriented transfer system, but also accounts for some semantic information

Reported 80-85% of output correct without post-editing.

Languages:

- English to German, Spanish
- German to English, French, Spanish, Danish
- French to English, Dutch
- Dutch to French

Langenscheidts T1 (Windows) based on METAL

- German to/from English; 320k word/phrase vocabulary

LMT

IBM

introduced in 1994; based on Logic Programming

Languages:

- English to/from German

LOGOS

Introduced in 1964

Multilingual Document Translation software

- English to German, French, Italian, Spanish, Portugese
- German to English, French, Italian
- French to German

ATLAS

Fujitsu, Ltd (Japan)

commercial since 1984; available in a PC version for Windows since 1994

base vocabulary: 130k words
24 subject fields totalling 1.2 million technical terms available as add-ons

Languages:

- Japanese to/from English

SYSTRAN

Systran (the company) was founded in 1968

- early development for US DoD (Russian->English for US Air Force)
- used by NASA for Apollo-Soyuz
- powers Altavista's Babelfish

38 language pairs developed, 16 commercially available

- English to French, German, Italian, Portugese, Spanish
- French to Dutch, English, German, Italian, Portugese, Spanish
- German to English, French
- Italian to English
- Spanish to English
- Portugese to English

`http://www.systransoft.com/`

English input:
`Hello, my name is Howard the Duck,`
`and I made a real turkey of a movie`
`some years ago.`

German translation:
`Vor hallo, ist mein Name Howard die Ente,`
`und ich bildete einen realen Truthahn von`
`einem Film einigen Jahren.`

English back-translation:
`Before hello, my name Howard is the duck,`
`and I formed a real truthahn of a film`
`unite for years.`

Problems:

- differing vocabularies for the different directions
  (turkey -> Truthahn, but Truthahn -> ?)
- reordering rules are available, but incorrectly applied

I have not sufficiently mastered German to allow my using it with impunity. My collection of fourteen-syllable German words is still incomplete. But I have just added to that collection a jewel -- a veritable jewel. I found it in a telegram from Linz, and it contains ninety-five letters:

```
Personaleinkommensteuerschatzungs-
kommissionsmitgliedsreisekosten-
rechnungserganzungsrevisionsfund *
```

If I could get a similar word engraved upon my tombstone I should sleep beneath it in peace.

-- Mark Twain, Vienna, 1899

*(Personal income tax estimation commission's member's travel expense bill reimbursement adjustment fund)

Other Systems

Eurotra:

- distributed, multinational translation project

GETA (Groupe pour d’Etudes pour la Traduction Automatique):

- ARIANE (developed since 1978)

PAHO (Pan-American Health Organization:

- SPANAM

A (dated) listing of MT and translation-aid software:
`www.fask.uni-mainz.de/ze/term/cat-pas.html`

Current Commercial Systems

Typically rule-based, with lexicons of 250-500k words/phrases and 1000 or more grammar rules.

Paradox?

In general, the older a system, the better it translates, regardless of its internals.

Why? Knowledge-acquisition bottleneck.

Problem with (direct-) transfer systems: for N languages, you need N^2 translation systems to cover arbitrary combinations.

Knowledge-Based MT

Humans know about the ’real’ world:

- The man saw the horse with the telescope.
- The man saw the girl with red hair.

KBMT attempts to encode such real-world knowledge to allow the MT system to disambiguate input and "understand" it.

Issues in KBMT

Expensive to acquire domain knowledge

- deciding which concepts to include
- relating concepts to each other
- what properties to encode

Deciding on granularity of knowledge:

- one-to-one mapping between concepts and lexical senses
- small number of conceptual primitives

Interlingual Systems

Warren Weaver, 1949:

> Think, by analogy, of individuals living in a series of tall closed towers, all erected over a common foundation. When they try to communicate with one another they shout back and forth, each from his own closed tower.... But when an individual goes down his tower, he finds himself in a great open basement, common to all the towers. Here he establishes easy and useful communication with the persons who have also descended from their towers. Thus it may be true that the way to translate from Chinese to Arabic, or from Russian to Portuguese, is not to attempt the direct route, shouting from tower to tower. Perhaps the way is to descend, from each language, down to the common base of human communication - the real but as yet undiscovered universal language - and then re-emerge by whatever particular route is convenient.

Advantages of an Interlingua:

- no N^2 blow-up as you add more languages
- language-independent intermediate rep

Disadvantages of an Interlingua:

- difficult to define language-indep. interlingua that includes *all* info needed for generation
- analysis and gernateion must be strictly separated

KBMT-89

An early knowledge-based system developed at CMU in 1987-89 for translating personal computer installation and maintenance manuals between English and Japanese.

- Lexical Functional Grammar parser to convert input into syntactic structures
- mapping rules to convert syntactic structure into semantic structure
- interactive disambiguation where needed
- more mapping rules to convert semantic structure into other language's syntactic structure
- generator to produce translated output

The full bidirectional translation system ran on six(!) workstations.

Techniques from KBMT-89 were incorporated into IBM Japan's SHALT-2 system (also English-Japanese translation).

KANT

KANT = Knowledge-based Accurate Natural-language Translation

An outgrowth of the KBMT-89 research, its main implementations are

- the Catalyst system for Caterpillar corporation
    - heavy equipment manuals
- the ESTRATO project for Union Electrica Fenosa
    - office documentation in electrical power domain

Next Lecture

Data-Driven MT Methods

- Statistical
- Example-Based

Combining the best of the best -- Multi-Engine MT