rijksuniversiteit
groningen

Machine Translation

Gideon Kotzé (g.j.kotze@rug.nl)

Rijksuniversiteit Groningen

28 September 2009

Important notice

- Content from slides 3, 4, 7, 8, 9, 10, 11 and 12 were taken from a Powerpoint slide show by Alon Lavie, called “Machine Translation Challenges and Language Divergences”
    - Obtainable at www.cs.cmu.edu/afs/cs/user/alavie/11-731/Spring-09/MT-Di
- slightly modified, mainly by including examples

Machine Translation: History

- **First MT device** was a mechanical dictionary, invented in 1933. Two patents (Artsrouni and Trojanskij).
- MT started in 1940's, one of the first conceived application of computers
- Promising "toy" demonstrations in the 1950's, failed miserably to scale up to "real" systems
- **ALPAC Report** (Automatic Language Processing Advisory Committee): MT recognized as an **extremely difficult**, "AI-complete" problem in early 1960's
- MT Revival started in earnest in 1980s (US, Japan)
- Field dominated by **rule-based approaches** (slow, costly)
- Economic incentive for developing MT systems for few language pairs (mostly European languages)

Machine Translation: Today

- **Age of Internet and Globalization** – great demand for MT:
  - Multiple official languages of UN, EU, Canada, etc.
  - Documentation dissemination for large manufacturers (Microsoft, IBM, Caterpillar)
- Economic incentive is still primarily within a **small number of language pairs**
- Some fairly good commercial products in the market for these language pairs
  - Still primarily a product of rule-based systems after many years of development
- Web-based (mostly free) MT services: Google, Babel Fish, others…

Google translate

Home | Text and Web | Translated Search

Translate text, webpage, or document

Enter text or a webpage URL, or upload a document.

Singer Marco Borsato is probably one of the most popular people in show business in the Netherlands.

English > Dutch swap Translate

Translation: English » Dutch

Zanger Marco Borsato is waarschijnlijk een van de meest populaire mensen in de showbusiness in Nederland.

Translate text, webpage, or document

Enter text or a webpage URL, or upload a document.

Monday morning the mail server for the complete university went down, due to data storage problems.

English > Dutch swap Translate

Translation: English » Dutch

Maandag ochtend de mail server voor de volledige universiteit ging, vanwege problemen met data-opslag.

Core challenges of MT

Ambiguity and Language Divergences:

- Human languages are highly ambiguous, and differently in different languages
- Ambiguity at all “levels”: lexical, syntactic, semantic, language-specific constructions and idioms
  - Lexical → level of **word**.
    - “bank” (financial institution or piece of land next to river?) --> called **homonyms/polysemes** (unrelated/related)
  - Syntactic → structure of the sentence.
    - “He saw the girl with the binoculars.”
      - Who’s got the binoculars?

rijksuniversiteit groningen

Core challenges of MT

- Semantic ambiguity: A word or phrase that can be interpreted in any number of ways
  - eg. “**You could do with** a new car”
  - Often seen in conventional and idiomatic expressions, etc.
- Idiomatic expressions are by nature ambiguous:
  - “He kicked the bucket.” -> “He died”, but also has a literal meaning.

Core challenges of MT

Amount of required knowledge:

- Translation equivalences for vast vocabularies (several 100k words and phrases)
- Syntactic knowledge (how to map syntax of one language to another), plus more complex language divergences (semantic differences, constructions and idioms, etc.)
- How do you acquire and construct a knowledge base that big that is (even mostly) correct and consistent?

Major sources of translation problems

- **Lexical Differences**:
  - Multiple possible translations for SL word (English: “bank”), or difficulties expressing SL word meaning in a single TL word (“ubuntu”)
- **Structural Differences**:
  - Syntax of SL is different than syntax of the TL: word order, sentence and constituent structure
    - “dat je **dat** moet doen” vs. “that you must do **that**”

Major sources of translation problems

- **Differences in Mappings of Syntax to Semantics**:
  - Meaning in TL is conveyed using a different syntactic structure than in the SL
- **Idioms and Constructions**
  - “the bigger they are, the harder they fall”
  - “keep a lid on this story”

How to tackle the core challenges

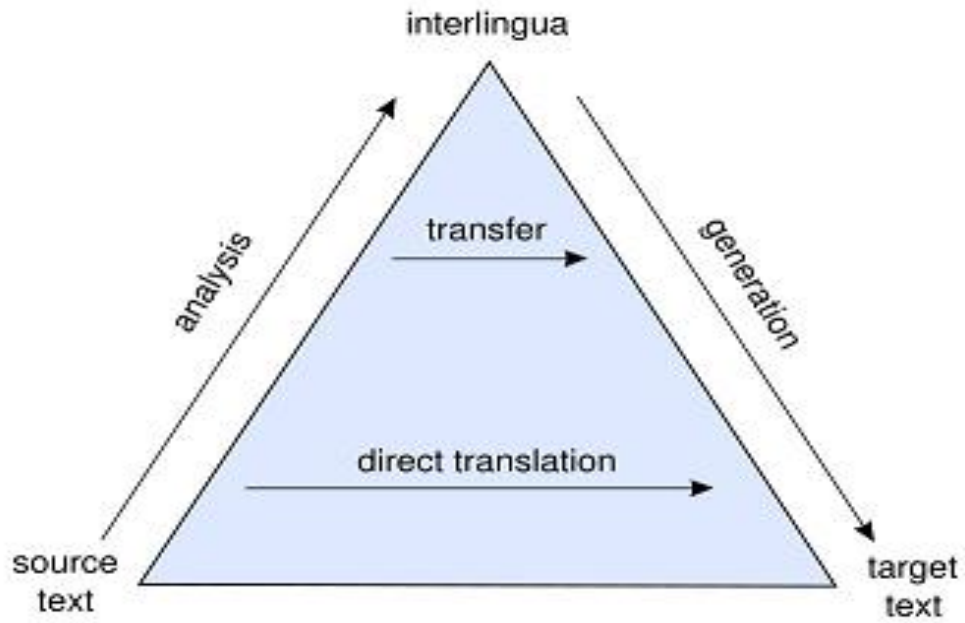
- **Manual Labor**: 1000s of person-years of human experts developing large word and phrase translation lexicons and translation rules (eg. Systran)
- **Lots of Parallel Data**: data-driven approaches for finding word and phrase correspondences automatically from large amounts of sentence-aligned parallel texts. Example: Statistical MT systems.
- **Learning Approaches**: learn translation rules automatically from small amounts of human translated and word-aligned data. Example: AVENUE's Statistical XFER approach.
- **Simplify the Problem**: build systems that are limited-domain or constrained in other ways. Examples: CATALYST, NESPOLE!

Main approaches

Rule-based

- direct (output: glosses, word-for-word → very bad!)
- transfer (Eurotra, Verbmobil, LOGON)
  - syntactic transfer
  - semantic transfer
  - interlingua (PIVOT, ATLAS II, Rosetta, DLT)
- direct easy but inaccurate, interlingua deemed too complex – most transfer systems somewhere inbetween
- Vauquois triangle

Vauquois triangle



Statistical Machine Translation

- Most successful approach today
- Examples: **Moses** (freely available), **Google**
- Translation = **machine learning problem**
- Models to induce alignments between n-grams
- Training data = large parallel corpora
- Advantages:
  - independent on language pair
  - good training = good results
  - fast and cheap to build
    - You can build your own MT system!

    (http://www.statmt.org/moses/)

Statistical Machine Translation

- **Disadvantages**:
  - system is a “black box”, difficult to detect easily correctable errors – no rules!
  - requires a lot of training data
- Uses Bayesian noisy channel model as a formalism to describe the problem:
- Translating from French to English: Best English sentence $\hat{E} = e_1,e_2,...,e_l$ given French sentence F is one where $P(E|F)$ is highest:

Statistical Machine Translation

$$\hat{E} = argmax_e P(E|F)$$
$$= argmax_E \frac{P(F|E)P(E)}{P(F)}$$

- Using and modifying Bayes' Rule, one can develop a basic model using n-grams
- **Word-based models** use the word as the basic unit – no syntactic knowledge at all.
- → Develop additional models for eg. **fertility** (number of times a words is repeated in translation before actually being translated), **reordering**, etc.
- Phrase-based models (PBSMT) has replaced word-based because of greater accuracy

Statistical Machine Translation

- PBSMT recognise that some words typically cluster together. Word alignment and phrase alignment probabilities give the clues. Important work:
  - Philipp Koehn, Franz Josef Och, Daniel Marcu: **Statistical Phrase-Based Translation** (2003)
- Many ways of modelling:
  - finite state transducers
  - synchronous context-free grammars
  - bracketing grammars
  - tree-adjoining grammars, etc.

Statistical Machine Translation

- Other important steps:
  - **Smoothing** (adjusting weights)
    - Idea: To account for sparse data, appearance of zero probabilities in test data (because of not appearing in training)
      - Different approaches: **Laplace smoothing, Good-Turing Discounting**, etc.
  - **Parameterization**
    - Assigning parameters to the models, using any one of a few machine learning methods

Statistical Machine Translation

- i.e. **generative models**, combining results of a language model and a translation model to obtain better estimates -> joint distributions
- or **discriminative models**, using features and assigning weights to these features according to their strength of contribution to a value
  - → conditional distributions

- **Parameter estimation**
  - → actually assigning the values to the parameters, using eg. **maximum likelihood estimation** for generative models

Statistical Machine Translation

- **Decoding**
  - Using the models, their parameters and values to actually produce a translation, seen as trying to produce the original form of a text that has been “corrupted”
- PBSMT still dominates today!
  - However, many recent attempts at **hybridization**.

Statistical Machine Translation

- **Why is PBSMT better than word-based models?**
  - Word-based models have problems with reordering that PBSMT addresses better by first splitting up the sentence into phrases.
  - By phrases are meant any arbitrary sequence of words, not the linguistic sense.
  - However, long-distance reordering presents problems.
  - In some cases, incorporating syntax helps, in others, not.
  - Many try to use more efficient grammars and language modelling.
  - Others opt for example-based approaches or more hybrid systems.

Example-based Machine Translation

- **EBMT** uses **analogy** by **matching** phrases with words extracted from corpora.
- Somewhere inbetween RBMT and SMT, not so clearly defined
  - Often, approaches from other paradigms also used
- No pure EBMT really exist...except maybe perhaps:
  - Lepage&Denoual (2005):

Example-based Machine Translation

"...it strictly does not make any use of variables, templates or patterns, does not have any explicit transfer component, and does not require preprocessing or training of the aligned examples. It only uses a specific operation, proportional analogy, that implicitly neutralises divergences between languages and captures lexical and syntactical variations along the paradigmatic and syntagmatic axes without explicitly decomposing sentences into fragments."

Example-based Machine Translation

- **Idea**: A:B :: C:D or “A is to B as C is to D”
- **Example from paper**
  - **English**:
  - **A**: Could you cash a traveler’s check?
  - **B**: I’d like to cash these traveler’s checks.
  - **C**: Could you open a window?
  - **D**: I’d like to open these windows.

Example-based Machine Translation

- **French**
  - **A**: Vous pouvez m'échanger un chèque de voyage?
  - **B**: Ces chéques de voyage, là, je peux les échanger?
  - **C**: Est-ce que vous pouvez m'ouvrir une fenêtre?
  - **D**: Est-ce que ces fenêtres, là, je peux les ouvrir?

- Translation using analogies between sentences of the same language and their translations.

Hybrid Machine Translation

- Many new systems propose a **hybrid approach**, trying to **combine** the advantages of other paradigms
  - Most have some statistical component
  - Most have some syntactic module
- Attempts at defining new **paradigms**:
  - “corpus-based”, “context-based”, etc.
- For the moment: interlingua and word-based models are forgotten

Hybrid Machine Translation: Metis-II

- Metis-II (2006) is a European MT effort for languages with little resources, using just:
  - dictionary
  - basic transfer system (few rules)
  - monolingual corpora
- Vandeghinste et al. (2006) add a few transfer rules which improves system performance.

Parse and Corpus-Based Machine Translation (PaCo-MT)

- PACO-MT attempts to **scale up** the approach in Metis-II.
  - very large corpora
  - combination of rule-based and data-driven approaches
    - hopes to achieve:
      - coverage of SMT (most rule-based systems have poor coverage)
      - transparency of rule-based systems (no “black box” syndrome)
  - based on belief that structural knowledge (syntactic, etc.) improves performance

Parse and Corpus-Based Machine Translation (PaCo-MT)

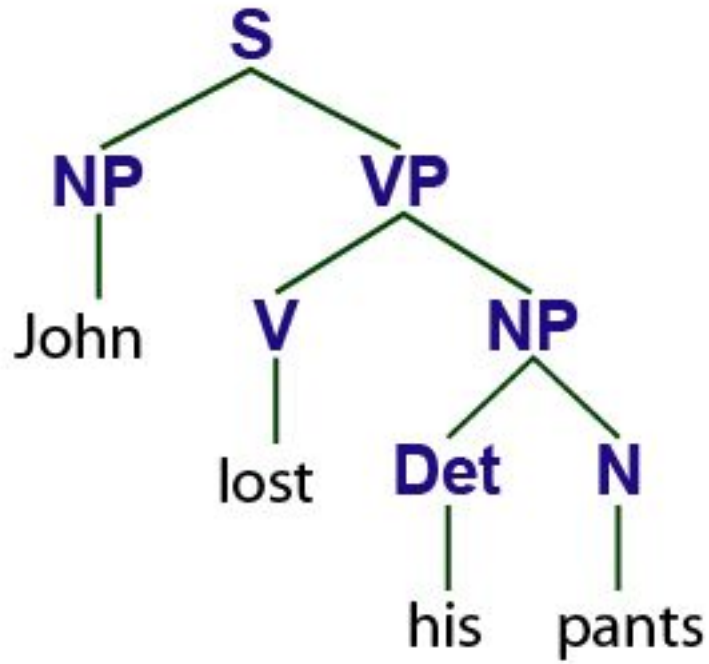
- Preparation and alignment of training data
  - Get bilingual and monolingual corpora
  - Sentence-align bilingual corpora
  - then word-align...
  - then parse everything...
  - then tree-align...
  - filter bad alignments

Parse and Corpus-Based Machine Translation (PaCo-MT)

- Next: **Automatic extraction of transfer rules** based on alignments
  - This is better than writing them yourself!
- Our partners in Leuven work on pipeline and a preliminary system
- Transfer on lexical level (words) and syntactic level (tree nodes)

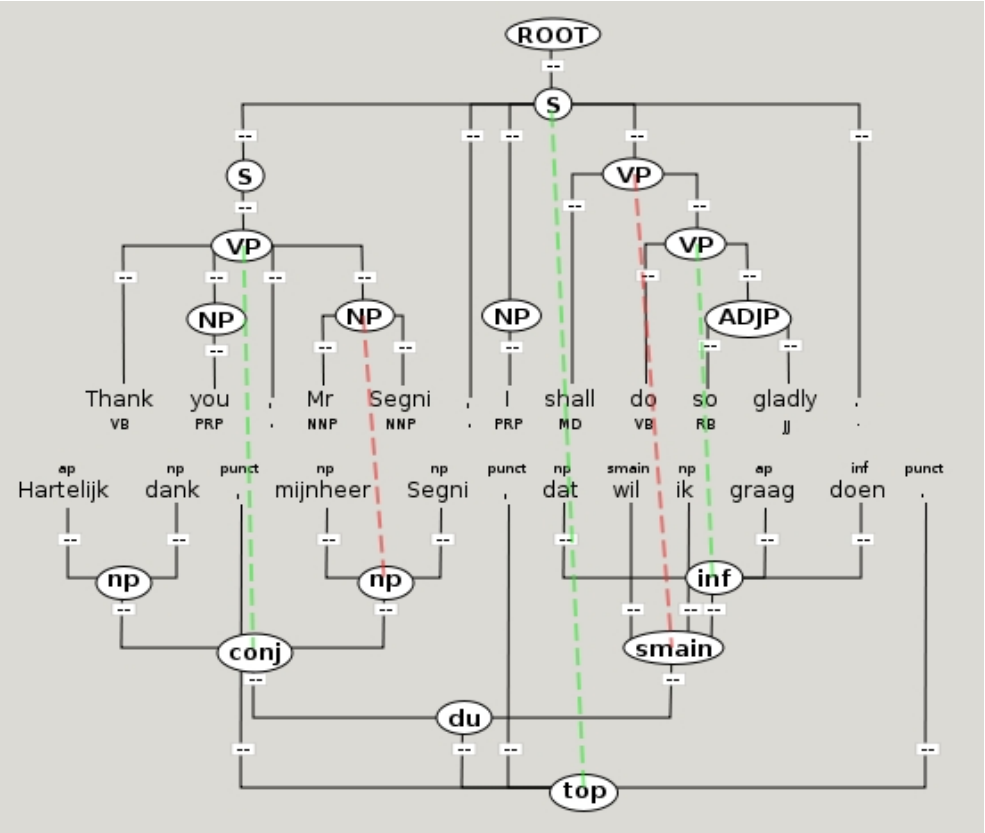
Parse and Corpus-Based Machine Translation (PaCo-MT)

Example of a syntactic tree:



Parse and Corpus-Based Machine Translation (PaCo-MT)

Example of an aligned tree pair:



Parse and Corpus-Based Machine Translation (PaCo-MT)

Example of a syntactic transfer rule:

```
{NP,1062753}
NP::NP [DNP NP] -> [NP PP]
(
(*score* 0.946640316205534)
(X2::Y1)
(X1::Y2)
)
```

Parse and Corpus-Based Machine Translation (PaCo-MT)

- Different methods of syntactic transfer:
  - **Tree-to-String** where only one side is parsed and using alignment information, rules are induced.
  - **Tree-to-Tree** where both sides are parsed.
    - →This will be our approach.
- No semantic transfer:
  - No semantic information available in training data.
  - Usually complicated, expensive.

Parse and Corpus-Based Machine Translation (PaCo-MT)

- Another important step is generation of a target side sentence
  - This is done via a lexicon and a model selecting the most likely words, constrained by the extracted rules.

Parse and Corpus-Based Machine Translation (PaCo-MT)

- Try to solve problems in terms of **accuracy**:
  - better word alignment, tree alignment and parsing, error mining and filtering of bad alignments
- ...in terms of **coverage**
    - eg. using k-best parses instead of just best parse
- ...decide on representation formats
  - phrase structure vs. dependency structure
  - file formats: Tiger-XML, or Penn, or Alpino...

Parse and Corpus-Based Machine Translation (PaCo-MT)

- ...also for rules themselves, eg. synchronous context free grammars, such as previous example
- In many recent approaches, trees are manipulated, often broken down into fragments and built up again, or flattened, etc.
- These (aligned) fragments correspond to rules.
- Using word and tree alignments, probabilities are derived for rules.

Parse and Corpus-Based Machine Translation (PaCo-MT)

- **Why syntactic transfer?**
  - → better than lexical transfer. Reasons:
  - Clue: PBSMT does better than word-based SMT
    - --> indicates that some words cluster together.
  - Syntactic transfer has no problem translating “blue sky” to the French “ciel blue” where words are **reordered**
    - Pure lexical transfer cannot do that.
  - **Extra information** is needed for the correct translation of “We ate cake yesterday” to “Wij hebben gisteren cake gegeten” instead of something like “Wij eten cake gisteren*”
    - only provided by morphosyntactic information

rijksuniversiteit groningen

Parse and Corpus-Based Machine Translation (PaCo-MT)

- Syntactic ambiguity is resolved by syntactically annotated sentences.
  - Eg. “The boy saw the girl with the binoculars.”
    - If “the girl with the binoculars” were annotated as part of the same noun phrase, then we would know that the girl has the binoculars.
- **Syntactic transfer does not always work**
  - For example, lexical ambiguity can only be resolved via statistical means, or using semantically annotated training data.
  - This is the main reason that hybrid systems exist: To make up for the shortcomings of the more “pure” approaches.

Thank you!

Questions, remarks, etc. welcome!