Human Language Technology:
Applications to Information Access

Lesson 5: Introduction to Machine Translation

October 27, 2016

EPFL Doctoral Course EE-724
Andrei Popescu-Belis
Idiap Research Institute, Martigny

Overcoming the cross-lingual barrier

- Part I of the HLT course dealt with the question: “how to find a needle in a haystack?”
  - but at least, we knew what a needle looks like
  - because experiments were in English

- But, in a globalized world … i.e. on the Web, relevant content may be in a language which differs from the query: so, can a system still find it?
  - if it finds it: how will you understand it?
  - options: query translation vs. document translation

➔ Need for **machine translation**

Plan of today's lesson (#5)

- Some uses and difficulties of machine translation (MT)
- Types of rules-based MT methods
  - direct | transfer | interlingua | example-based
- Principles of statistical machine translation (SMT)
  - phrase-based | hierarchical | neural [talk by L.Miculicich]
- Brief history of MT and landmark systems
- Measuring the quality of MT (evaluation)

Machine translation

- Computational method to translate from a *source language* into a *target language*
  - words < sentences < texts

- Two possible visions
  1. "fully-automatic high-quality MT" (FAHQMT)
     - replace human translators with machines
       - and even interpreters of spoken language (with ASR)
  2. "good applications for crummy MT"
     [Hovy & Church 1993]

Types of MT use

- Assimilation: user monitors large number of foreign texts
  - document routing / sorting
  - information extraction / summarization
  - cross-language information retrieval

- Dissemination: deliver texts in a foreign language to others
  - need for high-quality output
  - can be combined with human post-editing
    - CAT = computer-aided translation ≠ MT
    - specific tools or workbenches for CAT, e.g. “translation memories”

- Communication: real-time or delayed across languages

Role of the context of use

- Types of MT use (previous slide), but also:
  - Profiles of targeted users
    - SL and TL proficiency
    - available time
  - Types of source texts

→ Different requirements on MT models and expected quality levels

Difficulties of MT (1/2)

- Words do not have unique meanings + each meaning can have several translations = there are many options to choose from

  voler (FR) → *steal* or *fly* (EN)

  bank (EN) → *banque* or (*rive* or *berge* or *bord*) (FR)

- Multi-word expressions (idioms) cannot generally be translated by translating their components individually

  to kick the bucket (EN) → *casser sa pipe* (FR)

- Words are generally "inflected" in sentences: *voir* → *voient*

- Order of words in sentences vary greatly with the language

  Have you seen <u>him</u>? (EN) → *Hast du <u>ihn</u> gesehen?* (DE)

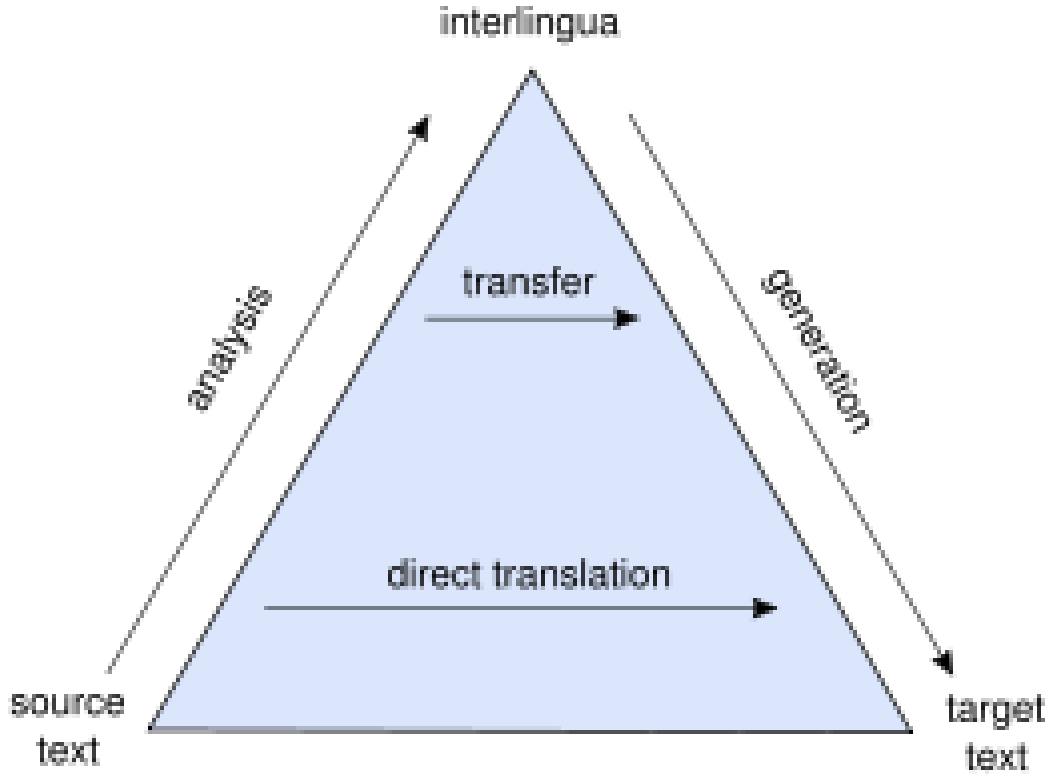
  → *<u>L</u>'as-tu vu?* (FR)

Difficulties of MT (2/2)

- Technical terms and compounds
- Pronouns: mismatches even between EN/FR
  - (FR) *il* / *elle* ↔ (EN) *he* / *she* / *it*
- Verb tenses: EN/FR mismatches
  - (FR) 'passé composé' / 'imparfait' ↔ (EN) 'simple past' / 'past perfect'
- Politeness-related phenomena
  - hard to guess, e.g. *you* ↔ *tu* / *vous*

- So it may seem that MT would require some form of "understanding" to address all these issues … or not?

Complexity of MT models: *Vauquois' triangle* a.k.a. MT pyramid



NB. The levels can be further subdivided

Machine translation models

- Rule-based MT
  - direct: word for word with local rewriting rules
  - transfer: analysis + transfer + synthesis
    - translation rules operate on a syntactic representation
  - interlingua: through a language-independent representation of the meaning (pivot or ontology)
- Corpus-based MT (data-driven or "empirical")
  - example-based (EBMT)
  - statistical (SMT): PBSMT, HMT, NMT
- Note: speech translation = ASR + MT (often SMT) + Synthesis

Direct MT

- No representation of meaning or syntactic structure
  - i.e. no grammar, no semantic resource, no ontology
- Knowledge is at the word level: “dictionaries”
- Dictionaries include, for each source word (and phrases)
  - lexical information (number, gender, etc.)
  - local syntactic constraints
  - possible translations with selection conditions and lexical information on translation
  - local reordering rules
- Translation: dictionary lookup | some disambiguation | search for translations | apply rules
- Robust, fast, flexible dictionaries

Deeper rule-based models

- Transfer-based MT
  - can operate on shallow syntactic representations, or more semantically-oriented ones (predicate/argument)
  - requires powerful and precise analysis components

- Interlingua-based MT
  - make real the dream of representing meaning
    - e.g. through an ontology such as UNL or CYC
  - adapted to limited domains with existing ontologies
  - seems appealing when many language pairs are needed, to reduce development costs from n^2 to n

Example-based MT

- Use a database of already translated examples to translate new sentences
  - cut the existing examples into meaningful chunks
  - determine the translations of chunks
- New sentence
  - cut it into chunks that are found in the database
  - generate new translation
- Can operate on linear chunks or on sub-trees
- Relationship to reasoning by analogy
- Connected to translation memories (CAT)

Statistical MT (Bayesian, generative)

- Translation as a noisy channel (W. Weaver)
  - source sentence *s* ↔ target sentence *t*
  - given *s*, what is the most likely translation *t*?

- Main idea
  - learn a translation model & a target-language model
  - decode source sentence: find most likely *t* given *s*

- P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, **19(2)**, 263-311. [Authors = team from IBM]

Formal definition

- Goal: given *s*, find *t* which maximizes P(*t*|*s*)

- Rewritten using Bayes's theorem:

$$\underset{t \in TL}{\operatorname{argmax}} P(t \mid s) = \underset{t \in TL}{\operatorname{argmax}} \left(P(s \mid t) \cdot P(t) \right)$$

translation model

language model (target)

Why not estimate and maximize $P(t|s)$ directly?

- Simplified answer: it is better to decompose the problem
  - a kind of "divide and conquer"
  - TM: how likely it is that a string is a translation of another string
  - LM: how likely it is that a string is well-formed

- Slightly less simplified answer
  - one can only approximate very roughly $P(t|s)$ for all sentences
    - this will often have non-zero probabilities on ill-formed strings
    - chances to find a well-formed string when maximizing $P(t|s)$ directly are close to zero
  - but, when maximizing $P(s|t)$
    - it doesn't matter if ill-formed strings receive non-zero probability
    - well-formedness is accounted for by the $P(t)$ term (language model)

1. The translation model

- Learned using a parallel corpus
  - i.e. many pairs of source and target sentences (translated by humans)
  - in SMT, it is often not important which one is the original sentences and which one is the human translation; parallel corpora often ignore this difference

- Goal : find a way to compute $P(s|t)$ given any s and t
  - starting with all (s, t) pairs of the corpus

- In other words, learn the parameters that will provide an estimate of $P(s|t)$ for a previously unseen (s, t) pair
  - idea: learn ***alignments*** between fragments of s and t, i.e. the parameters that represent how (groups of) words are related across languages

Of course, 1:1 alignment is quite infrequent.
Naturellement, un alignement 1 à 1 est très peu fréquent.

Word-based approach: use word "alignments" to compute probabilities of translation

$$P(s \mid t) = \sum_{a \in A(s,t)} P(s, a \mid t)$$

where $A(s, t)$ are all possible "alignments" of s and t

$$P(s, a \mid t) = \prod_{j=1}^{m} tr(s_j \mid t_{a_j})$$

where $tr(s_j \mid t_{a_j})$ is the translation probability of word t_{a_j} as word s_j, at positions j and a_j (= alignment variable)

Advanced translation models

1. Better than word-based: phrase-based models
   - alignments between “phrases” = groups of words, however not linguistically motivated phrases
   - phrase-based decoding: capture some lexical reordering, and translation of idiomatic expressions

2. Abstract transfer representations: hierarchical
   - useful to model reordering of words
   - using machine learning to learn how to parse
   - syntax can be used on source side, on target side, or both: tree-to-string | string-to-tree | tree-to-tree

2. Language modeling

- Probability of a given sequence of words in the target language, learned from a corpus
- Often n-gram based, e.g. trigram:

$$P(w_1,\ldots,w_m)=\prod_{i=1}^{m+2}P(w_i\mid w_{i-1},w_{i-2})$$

with provision for initial and final marks (≈words)

– noted generally <s> and </s>

3. Decoding

- Search for the best target sentence given the source sentence: $t_0 = \underset{t \in T}{\text{argmax}} \left(P(s \mid t) \cdot P(t) \right)$
- Greedy hill-climbing search
  - start with a word-for-word translation
  - trying various changes to improve likelihood
- Beam search decoding
  - examine source sentence from left to right
  - prune hypotheses to reduce search space

Some history of MT

- First attempts RU → EN in the 1950s
  - Weaver's code model, Georgetown experiment (IBM)
- ALPAC Report halts US funding in 1966
- Commercial success of SYSTRAN at end 1970s (EU)
- Rule-based systems in the 1980s, some interlingua ones

- Statistical MT made major progress since 1990s
  - related to progress in computing, modeling, metrics
  - PBSMT/HMT was the state-of-the-art until 2015 → neural MT

- Today: MT systems are still quite imperfect but widely used
  - individual or corporate use, Web-based, mobile devices

Examples of systems

- IBM Georgetown demonstration 1954
- METEO by TAUM 1981
- SYSTRAN company 1967
- Reverso by Promt and Softissimo 1997
- Metal / T1 / Comprendium 1985
- KANT and Catalyst by CMU for Caterpillar 1992
- UNL approach 1996
- Candide from IBM 1992
- Babelfish 1997
- Statistical tools 2000
  - GIZA++ aligner
  - Moses, Pharaoh, cdec
  - SRILM, IRSTLM
  - Europarl data
- Language Weaver 2002
- Google Translate 2006

Which MT method is better?

Consider the following example:

- Source sentence
  Les résultats d'études récentes le démontrent clairement : plus la prévention commence tôt, plus elle est efficace.
- Google translate (*PBSMT or NMT*)
  The results of recent studies show clearly: more prevention starts early, it is more effective.
- Systran box (*direct*)
  The results of recent studies show it clearly: the more the prevention starts early, the more it is effective.
- Systran PureNMT (*NMT, since October 2016*)
  The results of recent studies clearly demonstrate this: the more prevention starts early, the more effective it is.
- Metal / L&H T1 / Comprendium (*transfer*)
  The results of recent studies demonstrate it clearly: the earlier the prevention begins, the more efficient it|she is.

Measuring the quality of MT

- Exact quantification is difficult for non-humans
  - maybe as difficult as MT itself (with some reason)
  - more about it in Lesson 8
- MT errors are very varied in nature
  - have contributions to overall quality
- Perfect or unintelligible translations are easy to score (max / min), but what about intermediary ones?
- Two types of metrics
  - applied by humans
  - automatic ones: generally using a reference translation

Human-based metrics: subjective

- Generally rated per sentence, then averaged
- ***Fluency***: is output acceptable in the target language?
  - i.e., is it good French, English, etc.
  - monolingual judges are sufficient
- ***Adequacy***: does output convey same meaning as input?
  - requires bilingual judges or a reference translation
- ***Informativeness***
  - is it possible to answer a set of pre-defined questions using the translation, with the same accuracy as using the source?

- Also: reading time, post-editing time, HTER, Cloze test

Automatic reference-based metrics

- Compare a candidate translation to reference translations of the same input, prepared by professionals
- All reference translations are equally acceptable (no unique perfect translation), so use an average distance
- Examples
  - **BLEU**: compares n-gram overlap between the candidate translation and one or more reference translations
  - geometric mean of n-gram precision (n≤4) with brevity penalty
    - NIST version of BLEU considers information gain of n-grams
  - **Word Error Rate**: mWER, mPER
  - **METEOR**: harmonic mean of unigram precision and recall
    - accepts stemming and synonymy matching
- Extremely important for statistical MT as learning criterion

BLEU score (created by IBM for NIST in 2002)

$$BLEU = BP \cdot \exp(\sum_{n=1}^{N} w_n \log p_n)$$

$$BP = \min(1, \exp(1 - r/c))$$

(brevity penalty)

$$p_n = \frac{\sum_{C \in \{candidates\}} \sum_{ngram \in C} count_{in\_ref,\,bound}(ngram)}{\sum_{C \in \{candidates\}} \sum_{ngram \in C} count(ngram)}$$

r = length of reference translation
c= length of candidate translation

$count_{in\_ref,bound}$ () = number of n grams in common with reference(s), bound/clipped by maximum number of occurrences in reference

- 2-4 reference translations (concatenated)
- n-grams from 1 to *N* (often *N*=4), weighted (often 1/*N*)

Conclusion

- MT is one of the oldest fields of computer science and probably its first HLT application
- Looks simple: string to string conversion, but it is not (and it shouldn't be)
- Plans of the next lessons
  - language models: learning and testing LMs
  - translation models: learning based on text alignment
  - decoding (i.e. … translating)
  - evaluating translations

References

- Philipp Koehn, *Statistical Machine Translation*, Cambridge University Press, 2010
- Sergei Nirenburg, Harold L. Somers, and Yorick Wilks, Readings in Machine Translation, MIT Press, 2003 [includes some history]
- Christopher Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999 – Chapter 13, "Statistical Alignment and Machine Translation"
- Daniel Jurafsky and James H. Martin, *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Second Edition, Prentice-Hall, 2008 – Chapter 25, "Machine Translation"
- Proceedings of the Conferences of the Association for Computational Linguistics (ACL), of the Machine Translation Summits, of the **Workshop on Machine Translation (WMT)**, of EMNLP, EACL, EAMT, etc. – including *Computational Linguistics* and *Machine Translation* journals.