CmpE 590 Research Project

SYSTRAN
Systran History

• Founded in 1968
• Important breakthroughs
  • First hybrid MT system
  • First neural network MT system
Hybrid MT System

• Rule-based MT: the linguistic rules and general or domain specific resources
• The statistical analysis derived from monolingual corpus
Rule-based System Architecture
Rule-Based System Architecture

• Analysis Module
  • Converting source sentence into an intermediate structure with its syntactical information

• Transfer Module
  • transferring into another intermediate structure with syntactical information about target sentence
  • specific to each language pair
  • not re-usable

• Analysis and Generation modules are re-usable, 90% of the code
Analysis + Transfer + Generation
Paradigm

- Analysis (80%)
  - Language-family rules
  - Language-specific rules

- Transfer (10%)
  - Language-family transfer
  - Language-pair rules

- Generation (10%)
  - Language-family rules
  - Language-specific rules
Analysis Module

• language-family rules
• language specific rules
• the analysis of the ambiguity related to infinitive/finite verbs
  • unimportant when translating from German to English (“laufen”- “run”)
  • but required to describe a disambiguation rule to get the infinitive form
    (“laufen”-“correr” rather than “corren”) in Romance languages.
Transfer Module

- divided into different sub-modules based on language family
- language-pair rules
- transfer cost can be reduced thanks to the similarities between source and target languages
  
  e.g. translating from Spanish/ Italian/ Portuguese to Spanish/ Italian/ Portuguese does not need much effort
  
  the different languages like Spanish/ Italian/ Portuguese and German require more detailed transfer rules in order to reach correct word reordering.
XML-Based interface

• from the transfer module to the generation module the intermediate target structure is represented in XML-Based interface
• store all linguistic, typographical and structural information and markups that show which word are spellchecked and used dictionaries.
XML-based Interface
Generation Module

• developed in C++ and the object-oriented programming paradigm is implemented in terms of linguistic families.

• The abstract classes represent syntactic constituents like word, syntagm, clause and sentence.

• The methods are language independent operations:
  • updating the structures (checking inflection agreement between a noun and its adjectives, or a verb and its auxiliaries),
  • the synthesis of missing elements (subject personal pronouns are dropped in Spanish, but in German they should be synthesized)
  • reordering,
  • updating the target tree based on punctuation, typography.
Generation Module

• the more special classes inheriting from the base classes represents the features of language family.

• The new methods covering particular linguistic knowledge to language family such as Western, Romance, Germanic are added or the existing methods are overridden.

• For example, the Spanish / Italian / Portuguese classes additionally includes the methods about verbal auxiliaries and enclitic pronouns.
Statistical Post-Editing

• perform on the translation output produced by the rule-based system in

• The statistical information derived from monolingual corpora
  • adding new terms, entities and terminology to the current dictionaries,
  • resolving the ambiguities,
  • use of language model to select alternative translations, determiner choice, and reordering.
Statistical Post-Editing

• translate unknown words in the Systran RBS
• capture slight terminology changes but still preserving POS and meaning to improve fluency (e.g. politicians → politiciens vs the more commonly used “hommes politiques”),
• capture multiword expressions and locutions
• update
  • determiner (on political commitments → sur des engagements politiques vs. sur les engagements politiques)
  • preposition (across the Atlantic → à travers l’atlantique vs. de l’autre côté de l’atlantique),
  • pronoun, tense (should not be hidden → ne devraient... vs. ne doivent...), number and gender, and re-arrange word-ordering
## Statistical Post-Editing

<table>
<thead>
<tr>
<th>Source</th>
<th>SYSTRAN</th>
<th>SYSTRAN +SPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monetary policy can be used to stimulate an economy just as much as fiscal policy, if not more, in election years, which politicians will always want to do.</td>
<td>La politique monétaire peut être employée pour stimuler une économie juste comme beaucoup que la politique fiscale, sinon plus, en années d’élection, que les politiciens voudront toujours faire.</td>
<td>La politique monétaire peut être utilisée pour stimuler l’économie, tout comme la politique fiscale, pour ne pas dire plus, dans les années d’élection, que les hommes politiques voudront toujours faire.</td>
</tr>
<tr>
<td>Fortschritte der 12 Bewerberländer auf dem Weg zum Beitritt</td>
<td>Progress of the 12 applicant countries on the way to the entry</td>
<td>Progress of the 12 candidate countries along the road to accession</td>
</tr>
<tr>
<td>En una perspectiva a más largo plazo, habrá una moneda única en todo el continente.</td>
<td>In a perspective to more long term, there will be a unique currency in all the continent.</td>
<td>In a more long-term perspective, there will be a single currency for the whole continent.</td>
</tr>
</tbody>
</table>
Lexicographic Resources

• has been growing for over years

• The system consists of
  • the main dictionary
  • the transfer dictionary
  • the phrase-based dictionary
Main Dictionary

• Provide thousands of bilingual entries with all necessary information about source words and target words for each language pair.
• Obtained from the monosource / multilingual dictionaries carefully prepared by lexicographers.
• For instance, French-to-English bilingual main dictionary is derived from the multitarget French source dictionary.
• Collect only grammatical words: prepositions, pronouns, particles, conjunctions, and very fundamental verbs: copulas, auxiliaries, linking verbs and verbs used commonly in periphrastic idioms and homographic words
Transfer Dictionary

• Built based on Intuitive Coding Technology
• Intuitive Coding Technology allows user to easily enter information to bilingual dictionaries.
• However, the inflection paradigms and homographs are guessed by the internal lexicographic rules and heuristics about morpho-syntax.
• Includefull words: nouns, verbs, adjectives and idiomatic sequences.
Phrase-based Dictionary

• Written by the statistical information about phrase-pairs extracted from parallel corpora.

• By the way, the existing rule-based, general-purpose MT system can be adapted into a given domain.

• Learning phrasal entries for rule-based system:
  • capture domain context to disambiguate the translation.
  • word-to-word translation is not of use in phrase translation
  • Strong collocations may resolve the syntactic ambiguity in the source sentence
Phrase-based Dictionary

• limited to phrases having the same syntactic structure
  e.g. noun phrases should be translated as noun phrases

• In the extraction process,
  • Words and phrases in each corpus are aligned
  • For each candidate phrase pair, some statistical features like frequency in the corpus, lexical weights in both directions, the lemma counts are calculated
  • Only the most frequent or best aligned (according to lexical weight) translation per source phrase is elected
  • Each extracted phrase-pair which delivers translation score calculated based on BLUE metrics below the threshold is pruned out
  • only remaining ones are added to current dictionaries in order to prevent to damage the rule-based system
Phrase-based Dictionary

Parallel text

Word Alignment

Phrase Alignment

Linguistic Filtering & Coding

Cumulative & Additional Filtering

Bilingual Dictionary
Neural Machine Translation

• deep learning approach to deliver high performance in several large-scale translation
• Possible to directly model the relationship between an input text in a source language and its translation in a target language in neural MT systems
• But very expensive in terms of computation and training duration.
Neural Machine Translation

• Systran NMT has been developed based on the open source project seq2seq-attn maintained by Harvard NLP group.
• The first generation works for generic translations, and covers 12 languages for 32 language pairs.
seq2seq-attn

- seq2seq-attn project provides various features
  - training with bidirectional encoders
  - pre-trained word embeddings
  - handling unknown words in decoding by substituting them with themselves or looking up their translation from an external resource
  - switching between CPU and GPU for both training and decoding.
Neural Machine Translation

• Integration of several features into the existing framework
• In tokenization, standard token separators (spaces, tabs, etc.) and language dependent linguistic rules are utilized
• Support for an arbitrary number of discrete word features as additional inputs to encoder and decoder
  • The features are represented in continuous and normalized vectors
  • Concatenated the vectors to word embeddings.
  • Supporting additional features on the target part is implemented by generating feature at time t+1 for the word generated at time t.
Neural Machine Translation

• The internal Named Entity modules previously developed for RMBT and SMT systems.

• Re-implement guided alignment strategy which guides attention mechanism in a NMT like IBM model 4 Viterbi alignments to handle placeholder substitutions and unknown words.

• Politeness feature is added to each source sentence in training very meaningful when translating from a language like English to a language including politeness expressions like Korean
Neural Machine Translation

• In the customization process,
  • Incrementally adapted into a specific domain by running additional training epochs over newly available in-domain data.
  • Even in limited in-domain data, the system demonstrates improved results.
• In post-editing,
  • monolingual and multi-source Neural Post Editing systems trained on the same data
  • Even multi-source NPE deliver similar performance to NMT after parameters are converged they are better than SPE.
Neural Machine Translation

• Training resources for each language pair:
  • a baseline corpus (1 million sentence) for day-scale experiments
  • a medium corpus (2-5 million sentences) for week-scale experiments
  • a very large corpora (more than 10 million sentences)