Overview

Introduction
  History

Multi-engine architectures
  Approach 1
  Result
  Approach 2
  Results

Conclusion
significant progress in coverage and quality of machine translation

focused translation of English
  1. simple morphological structure
  2. abundance of monolingual training data

→ lack of linguistic sophistication of the language models
trained on massive amounts of data
→ capture implicit knowledge in co-occurrence statistics (serves as shallow replacement for world knowledge):

1. required for ambiguity resolution
2. insertion of missing information, required to produce well-formed text in target language
MT Systems

► emphasis on MT systems for frequently used language pairs:
MT Systems

emphasis on MT systems for frequently used language pairs:

1. wealth of linguistic knowledge about the languages
MT Systems

► emphasis on MT systems for frequently used language pairs:
  1. wealth of linguistic knowledge about the languages
  2. complete mechanisms for morphological and syntactic analysis and generation
MT Systems

- emphasis on MT systems for frequently used language pairs:
  1. wealth of linguistic knowledge about the languages
  2. complete mechanisms for morphological and syntactic analysis and generation
  3. large number of bilingual lexical entries
The challenge

Combine different types of knowledge into integrated systems:
→ linguistic knowledge contained in the rules of one or several conventional MT systems
→ non-linguistic knowledge extracted from large amount of text
EuroMatrix Project

- exploring combinations of rule-based and statistical knowledge sources
- **Approach**: integration of existing rule-based MT systems into multi-engine architecture
The approach

Multi-engine architectures

1. select the best output from number of systems, leaving the individual hypothesis
2. try to recombine the best parts from multiple hypotheses into a new utterance, better than the best of given candidates

Requirements: finding correspondences between alternative renderings of a source-language expression
Difficulties: different word order, errors in output → difficulties in alignment identification
The approach and SMT Decoder

- involve word alignment between MT output and source text
- search procedure for naturally sounding combinations of highly probable partial translations - SMT decoder

The steps:

1. transform information from MT output into suitable input for decoder
2. find good relative weights for various phrase table entries

The risk: linguistically well-formed constructs from one of the rule-based engines will deteriorate in final decoding step
Setup

1. set of six rule-based MT engines (Systran, Lucy Soft (METAL, Comprendium), Personal Translator, Translate (LMT), MorphoLogic, ProMT, Tilde, etc.)
2. generated phrase tables from training data
3. improved alignment quality (source and output text) - using modified version of GIZA++
Results

Hybrid Machine Translation Architectures

Nora Kumpikova

Introduction
History
Multi-engine architectures
Approach 1
Result
Approach 2
Results
Conclusion
Results

→ Testing is done in domain, English as the target language:
  ▶ the statistical approaches can adapt to the domain’s typical expressions
  ▶ the best statistical systems are better than the best RBMT system in sentence ranking and much better in constituent ranking
→ For tests in a different domain:
  ▶ the rule-based systems are better than SMTs in sentence ranking
  ▶ the rule-based systems have only a very slight advantage in constituent ranking
→ Under both scenarios:
  ▶ the hybrid combination behaves similar to the SMT system (can obtain a slight improvement from the larger lexicon)
The approach

**Goal:** Feeding SMT phrases into rule-based MT system

**Motivation**
Investigate how automatically extracted lexical knowledge can be used to increase coverage of rule-based MT system. → rule-based system controls of the transition process

**Advantages:** well-formed syntactic structures generated by linguistic rules cannot be broken up by the SMT components

**Disadvantages of rule-based systems:** lack mechanisms for ruling out implausible results, cannot cope with errors in lexicon due to misalignments/examples that fail to generalize, expressions that depend on the context
Processing lexical entries

- DFKI - extraction and manual validation of additional lexical entries for relevant technical fields
- European Patent Office (EPO) - translation facility for patent documents
- Parallel texts processed through statistical alignment algorithms (GIZA++)
- Texts linguistically enriched by PoS-tags and lemma information
- Combine representations → obtain a set of candidates for the lexicon

→ Automatic creation of lexical entries: manually inspect/correct results to prevent errors
Results

Create translation for:

1. English ↔ German: 40 million sentences processed → 2.3 million candidates EN-DE term pairs identified
2. English ↔ Spanish: 10 million sentences processed → 0.8 million candidates EN-ES term pairs identified

▶ 90% of extracted entries are pairs of noun phrases (typically multi-word expressions)
▶ often English MWE correspond to one long German word
Both approaches show promising improvements in MT quality

too early to give meaningful comparative evaluations

does not discuss that the output of RBMT engines often sounds less natural/fluent in comparison to SMT results

difficulty to deal with disambiguation and picking most natural expressions in target language

suggested by authors: a deeper integration of rule-based linguistic knowledge with corpus-based evidence will address both issues in one integrated system