Machine Translation and Cross-language Information Retrieval

[DAT640] Information Retrieval and Text Mining

Petra Galuscakova
University of Stavanger

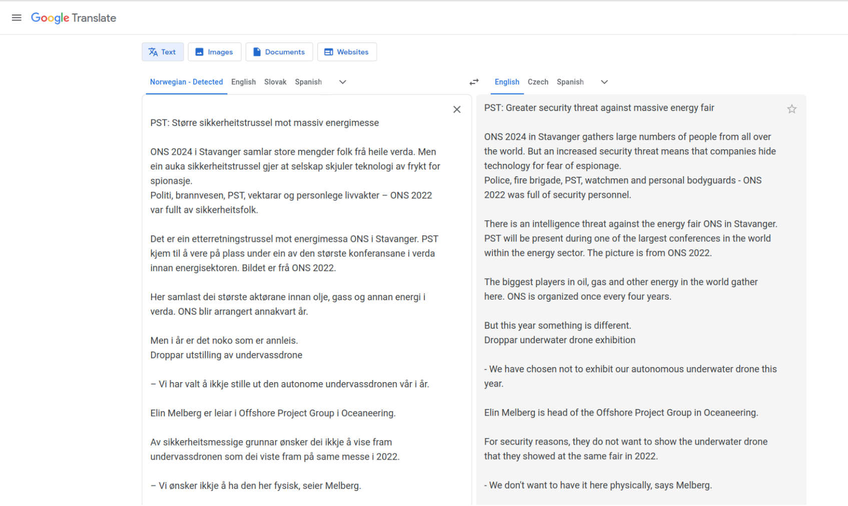
November 12, 2024

In this module

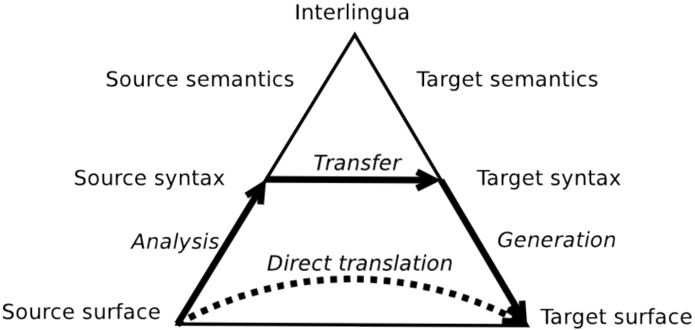
1. Machine Translation
2. Cross-language Information Retrieval

Machine Translation

What is MT?


Google Translate
Text
Images
Documents
Websites
Norwegian - Detected
English
Slovak
Spanish
English
Czech
Spanish
PST: Større sikkerheitstrussel mot massiv energimesse
ONS 2024 i Stavanger samlar store mengder folk frå heile verda. Men ein auka sikkerheitstrussel gjer at selskap skjuler teknologi av frykt for spionasje.
Politi, brannvesen, PST, vektarar og personlege livvakter – ONS 2022 var fullt av sikkerheitsfolk.
Det er ein etterretningstrussel mot energimessa ONS i Stavanger. PST kjem til å vere på plass under ein av den største konferansane i verda innan energisektoren. Bildet er frå ONS 2022.
Her samlast dei største aktørane innan olje, gass og annan energi i verda. ONS blir arrangert annakvart år.
Men i år er det noko som er annleis.
Droppar utstilling av undervassdrone
– Vi har valt å ikkje stille ut den autonome undervassdronen vår i år.
Elin Melberg er leiar i Offshore Project Group i Oceaneering.
Av sikkerheitsmessige grunnar ønsker dei ikkje å vise fram undervassdronen som dei viste fram på same messe i 2022.
– Vi ønsker ikkje å ha den her fysisk, seier Melberg.
PST: Greater security threat against massive energy fair
ONS 2024 in Stavanger gathers large numbers of people from all over the world. But an increased security threat means that companies hide technology for fear of espionage.
Police, fire brigade, PST, watchmen and personal bodyguards - ONS 2022 was full of security personnel.
There is an intelligence threat against the energy fair ONS in Stavanger. PST will be present during one of the largest conferences in the world within the energy sector. The picture is from ONS 2022.
The biggest players in oil, gas and other energy in the world gather here. ONS is organized once every four years.
But this year something is different.
Droppar underwater drone exhibition
- We have chosen not to exhibit our autonomous underwater drone this year.
Elin Melberg is head of the Offshore Project Group in Oceaneering.
For security reasons, they do not want to show the underwater drone that they showed at the same fair in 2022.
- We don't want to have it here physically, says Melberg.

Types of MT Systems (Vauquois Triangle)



- Approaches to MT can be categorized by whether they work directly with surface words or whether they utilize some (linguistic) abstraction.
- Some MT systems disregard any linguistic information and treat all words as unrelated, indivisible units.
- Other systems perform linguistic analysis on the source side and then do transfer – either to some abstract representation or directly to target-side surface words.

Levels of Language Description recap.

- Phonetics [Sounds; (nearly) language independent]
- Phonology [Sound patterns, language dependent abstraction over sound]
- Morphology [Word structure]
- Syntax [Sentence structure]
- Semantics [Literal meaning]
- Pragmatics [Meaning in context]

Types of MT Systems cont.

- Another possible distinction is how the systems are ”trained”
- Rule-based systems: human experts would manually develop rules to describe the analysis, transfer or generation for a particular language pair.
- Statistical systems require data and utilize statistical models or machine learning to capture the knowledge required for translation.
- Neural models use encoder-decoder architecture.
- Generative models use decoder-only LLMs.

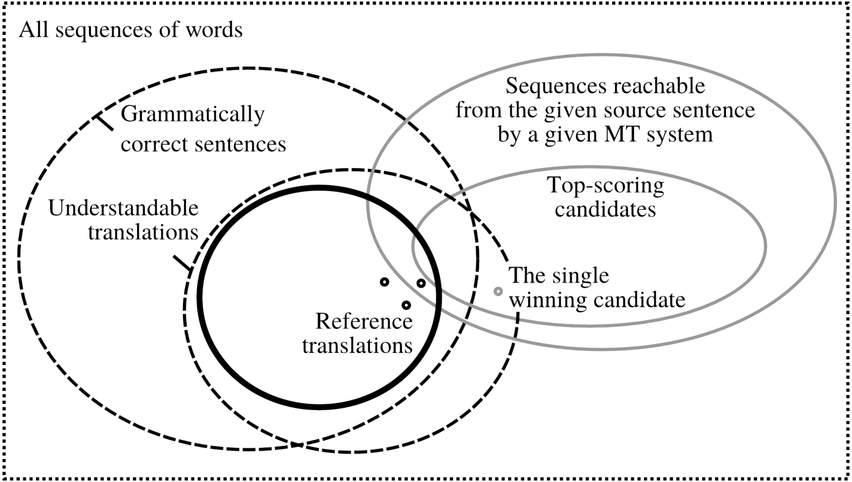
Evaluation of MT Systems

- We restrict the task of MT to the following conditions:
    - No writers' ambitions, we prefer literal translation.
    - No attempt at handling cultural differences.
- Evaluation might be done manually or automatically.
- Manual evaluation:
    - It is often done by relative ranking of full sentences by several MT systems.
    - Adequacy, fluency and comprehension of whole sentences or its constituents might be also directly judged.
- Automatic evaluation is fast and cheap, deterministic, replicable and it allows automatic model optimization, but less corresponding to real preferences.
- BLEU (Bilingual evaluation understudy) remains the most popular metric for automatic evaluation of MT output quality.

Automatic Evaluation of MT Systems

- BLEU considers sequences of words: the amount of overlap of n-grams between the candidate translation and the reference (more specifically unigrams, bigrams, trigrams and 4-grams, in the standard formulation).
- The formal definition is as follows: $BLEU = \mathrm{BP} \cdot \exp \sum_{i=1}^{n} (\lambda_i \log p_i)$
- Where (almost always) $\lambda_i = 1/n$ and n = 4, p_i stand for i-gram precision, i.e. the number of i-grams in the candidate translation which are confirmed by the reference.
- Each reference n-gram can be used to confirm the candidate n-gram only once (clipping), making it impossible to game BLEU by producing many occurrences of a single common word (such as "the").
- BP stands for brevity penalty. Since BLEU is a kind of precision, short outputs (which only contain words that the system is sure about) would score highly without BP.

Issue with Automatic Evaluation[1]



[1] https://mttalks.ufal.ms.mff.cuni.cz/index.php/Automatic_MT_Evaluation

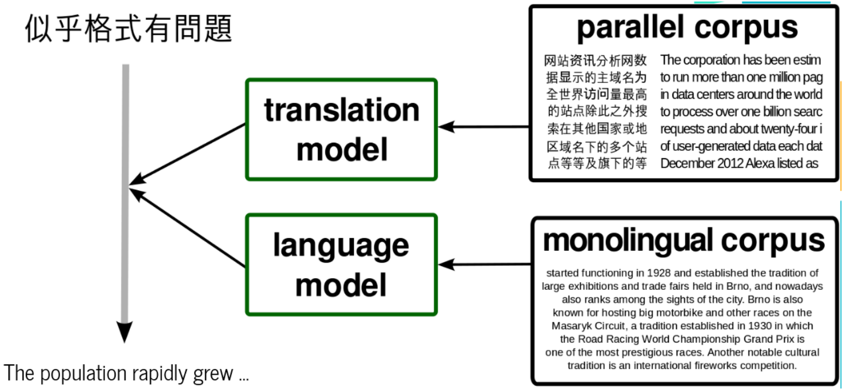
Rule-based MT

- Works using the rules specified by human experts.
- Can work well for closely related languages (Norwegian and Swedish, Norwegian and Danish, Czech and Slovak, Spanish and Portuguese, ...)
- Typically works in several stages:
  - 1) Getting basic part-of-speech information of each source word (e.g. a = indef.article; girl = noun; eats = verb; an = indef.article; apple = noun)
  - 2) Getting syntactic information about the verb (e.g. "to eat": NP-eat-NP; here: eat – Present Simple, 3rd Person Singular, Active Voice)
  - 3) Parsing the source sentence (e.g. NP an apple = the object of eat)
  - 4) translate English words into German (e.g. a (category = indef.article) => ein (category = indef.article), girl (category = noun) => Mädchen (category = noun), eat (category = verb) => essen (category = verb), an (category = indef.article) => ein (category = indef.article), apple (category = noun) => Apfel (category = noun))
  - 5) Mapping dictionary entries into appropriate inflected forms (final generation): A girl eats an apple. => Ein Mädchen isst einen Apfel

Statistical machine translation (SMT)

- Given a source (foreign) language sentence $f_1^J = f_1 ... f_j ... f_J$ produce a target language (English) sentence $e_1^I = e_1 ... e_j ... e_I$
- Among all possible target language sentences, choose the sentence with the highest probability: $\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{argmax} \quad p(e_1^I | f_1^J)$
- After applying Bayes law: $\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{argmax} \quad p(f_1^J | e_1^I) p(e_1^I)$
- $p(f_1^J | e_1^I)$ is a translation model
- $p(e_1^I)$ is a language model
- The most successful SMT approach was phrase-based machine translation (PBMT): count co-occurences of phrase pairs $(\hat{f}, \hat{e})$

SMT Models



似乎格式有問題
translation model
language model
parallel corpus
网站资讯分析网数
据显示的主域名为
全世界访问量最高
的站点除此之外搜
索在其他国家或地
区域名下的多个站
点等等及旗下的等
The corporation has been estim
to run more than one million pag
in data centers around the world
to process over one billion searc
requests and about twenty-four i
of user-generated data each dat
December 2012 Alexa listed as
monolingual corpus
started functioning in 1928 and established the tradition of large exhibitions and trade fairs held in Brno, and nowadays also ranks among the sights of the city. Brno is also known for hosting big motorbike and other races on the Masaryk Circuit, a tradition established in 1930 in which the Road Racing World Championship Grand Prix is one of the most prestigious races. Another notable cultural tradition is an international fireworks competition.
The population rapidly grew ...

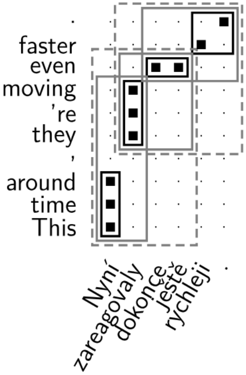
Traditional SMT Pipeline

- Training the Translation Model:
  - Find relevant parallel texts.
  - Align at the level of sentences.
  - Align at the level of words.
  - Extract translation units, with scores (co-oc. stats.).
- Tuning = Actual training in the ML sense
  - Identify TM/LM/other model component weights.
- Translation:
  - Decompose input into known units.
  - Search for best combinations of units.

Translation Model Example

in europa ||| in europe ||| 0.829007 0.207955 0.801493 0.492402
europas ||| in europe ||| 0.0251019 0.066211 0.0342506 0.0079563
in der europaeischen union ||| in europe ||| 0.018451 0.00100126 0.0319584 0.0196869
in europa , ||| in europe ||| 0.011371 0.207955 0.207843 0.492402
europaeischen ||| in europe ||| 0.00686548 0.0754338 0.000863791 0.046128
im europaeischen ||| in europe ||| 0.00579275 0.00914601 0.0241287 0.0162482
fuer europa ||| in europe ||| 0.00493456 0.0132369 0.0372168 0.0511473
in europa zu ||| in europe ||| 0.00429092 0.207955 0.714286 0.492402
an europa ||| in europe ||| 0.00386183 0.0114416 0.352941 0.118441
der europaeischen ||| in europe ||| 0.00343274 0.00141532 0.00099583 0.000512159

Translation Alignments[2]



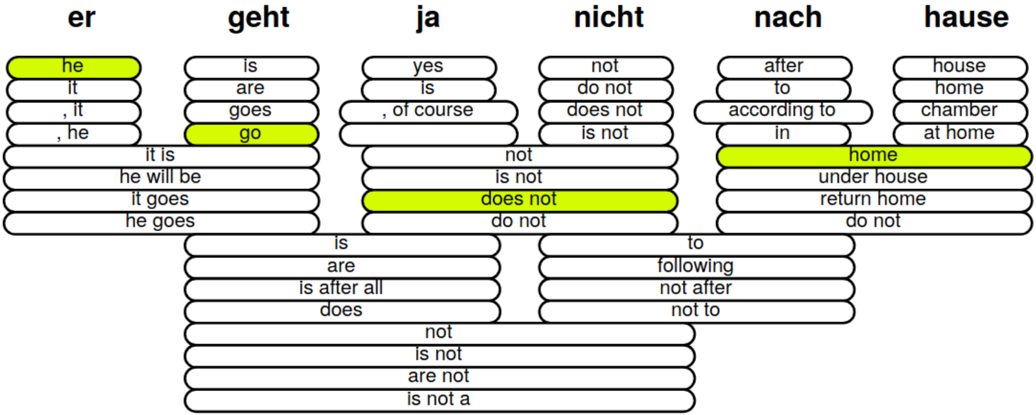
- Translation in PBMT: 1) Extract all phrases (up to max-phrase-len) and 2) score them

[2] https://ufal.mff.cuni.cz/~bojar/courses/npfl087/1920/05-pbmt.pdf

Translation Alignments cont.

- Given parallel training corpus, phrases in PBMT are extracted and (consistent with the word alignments – might be long and short, overlapping in all ways) and then scored
- Alignments:
  - This time around = Nyní
  - they 're moving = zareagovaly
  - even = dokonce ještě
  - ... = ...
- Phrases:
  - This time around, they 're moving = Nyní zareagovaly
  - even faster = dokonce ještě rychleji
  - ... = ...

Translation Decoding[3]



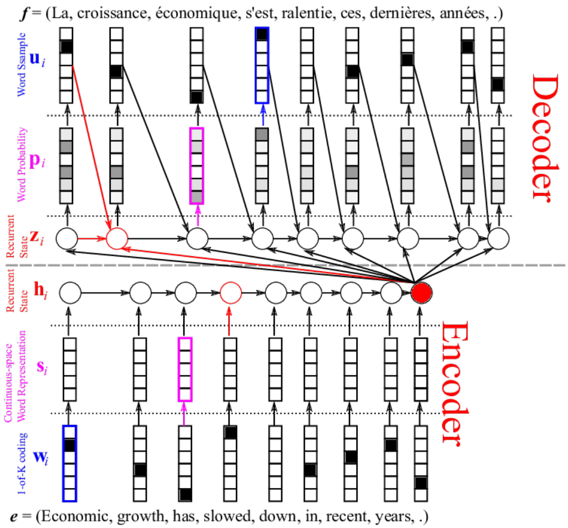
- It is done using Beam search

[3] https://ufal.mff.cuni.cz/∼bojar/courses/npfl087/1920/05-pbmt.pdf

Neural Machine Translation (NMT)

- First NMT systems were based on sequence-to-sequence encoder-decoder model
- Once processing reaches the end of the input sentence the hidden state encodes its meaning (encoding phase).
- Then this hidden state is used to produce the translation in the decoder phase.
- In practice, the proposed models works reasonable well for short sentences (up to, say, 10–15 words), but fails for long sentences.
- Later, this was improved by the attention mechanism.

NMT using RNN



NMT using RNN

- In 2022-23 the dominant architecture was encoder-decoder models which used transformers.
- These models are especially especially useful for low-resource languages, where large parallel datasets do not exist.
- Later, it was found out that 'GPT systems can produce highly fluent and competitive translation outputs even in the zero-shot setting especially for the high-resource language translations' (Hendy et al., 2023).
- GPT systems are still behind for the low resource languages.
- GPT systems might be used using zero-shot (Translate this sentence from [source language] to [target language], Source: ... Target:) or few-shot approaches.

Translation Quality Comparison[4]

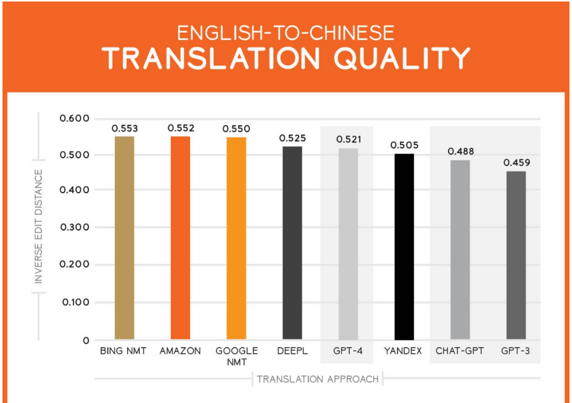


Figure 1. Comparison of automated translation quality between GPT models and the five major Neural MT engines based on the inverse edit distance using multiple references for the English-to-Chinese language pair.

[4] https://www.lionbridge.com/blog/translation-localization/machine-translation-a-generative-ai-model-outperformed-a-neural-machine-translation-engine/

Translation Quality Comparison[5]

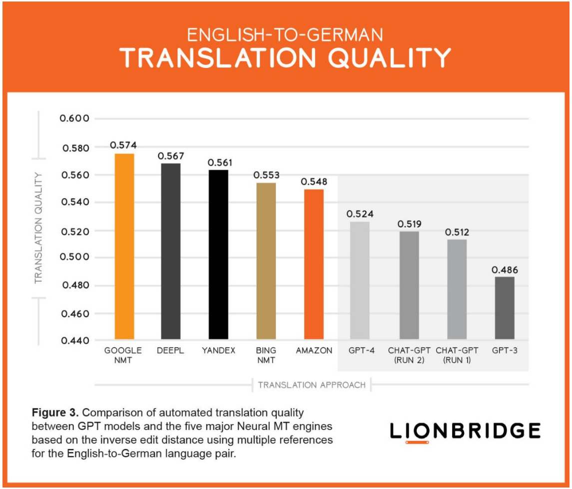


Figure 3. Comparison of automated translation quality between GPT models and the five major Neural MT engines based on the inverse edit distance using multiple references for the English-to-German language pair.

[5] https://www.lionbridge.com/blog/translation-localization/machine-translation-a-generative-ai-model-outperformed-a-neural-machine-translation-engine/

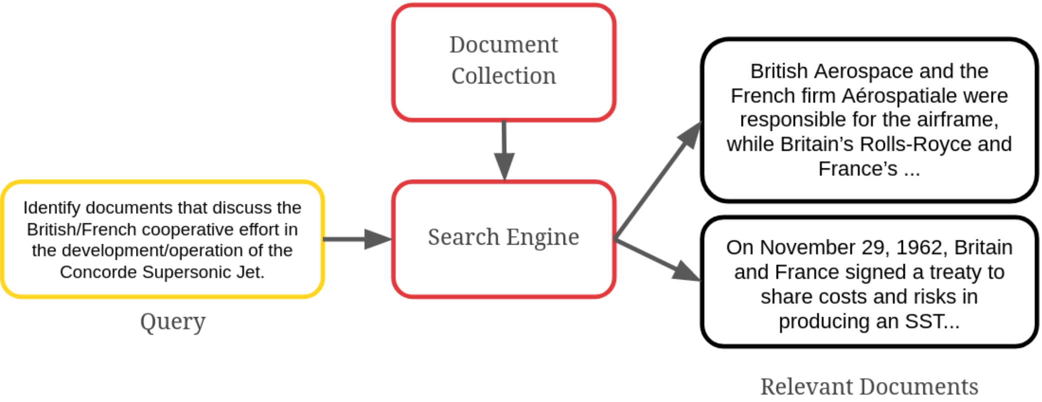
Translation Quality Comparison[6]

Figure 2. Comparison of automated translation quality between GPT models and the five major Neural MT engines based on the inverse edit distance using multiple references for the English-to-Spanish language pair.

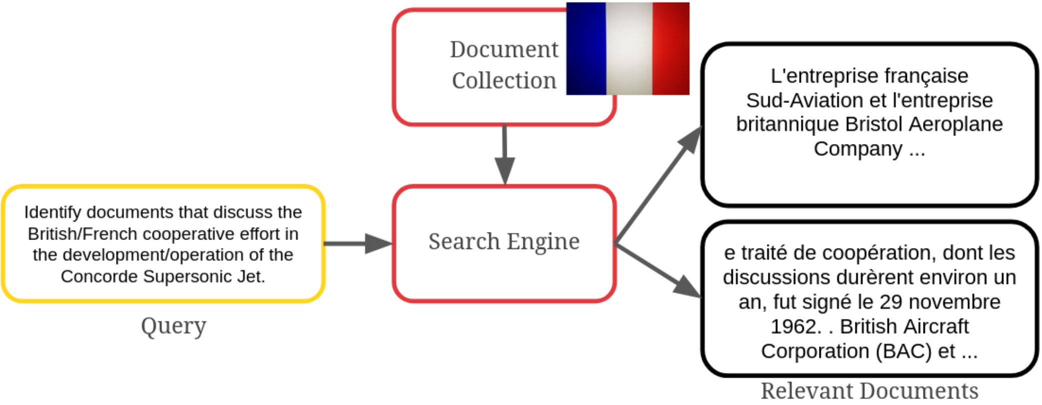
[6]https://www.lionbridge.com/blog/translation-localization/machine-translation-a-generative-ai-model-outperformed-a-neural-machine-translation-engine/

Cross-language Information Retrieval

Information Retrieval setup



Cross-language Information Retrieval



Cross-language Information Retrieval cont.

- The goal of Cross-Language Information Retrieval (CLIR) is to build search engines that use a query expressed in one language (e.g., English) to find content that is expressed in some other language (e.g., French)
- Two key assumptions shape the usual view of ranked retrieval: (1) that the searcher can choose words for their query that might appear in the documents that they wish to see, and (2) that ranking retrieved documents will suffice because the searcher will be able to recognize those which they wished to find
- When the documents to be searched are in a language not known by the searcher neither assumption is true
- CLIR is closely linked with Machine Translation; what we call MT is the use of translation technology to render documents readable, whereas CLIR is the use of translation technology to render documents searchable

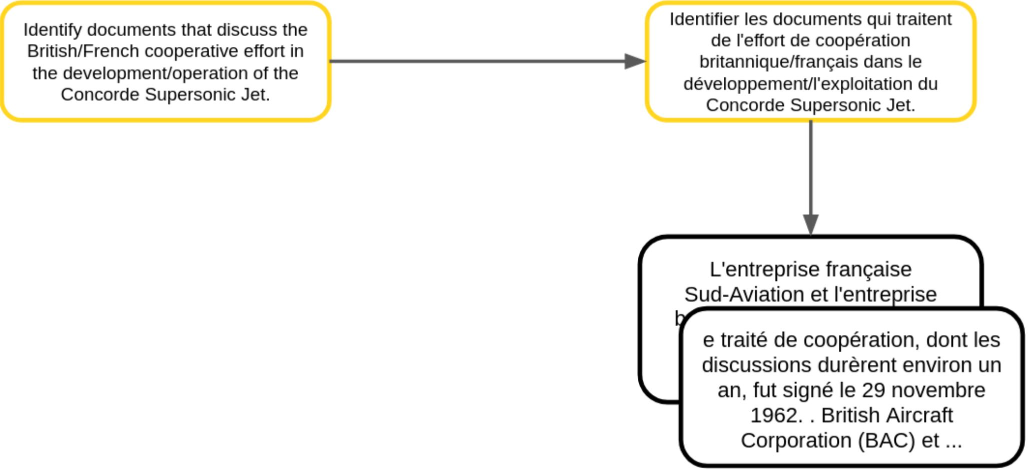
Cross-language Information Retrieval use cases

- Two main use cases for CLIR:
  - The user lacks proficiency in the language of the documents
  - The user can understand the documents, but prefers to use a different language
- In Web search, there is the natural emergence of one or more "lingua franca" languages such as English or Chinese which can be used to query the Web
- CLIR applications can be expected to be particularly important in regions where multiple languages are frequently use (English/French in Canada, Dutch/French/German in Belgium, Spanish/Catalan/Basque in Spain)

Exercise

E16-1 Design an CLIR System using MT.

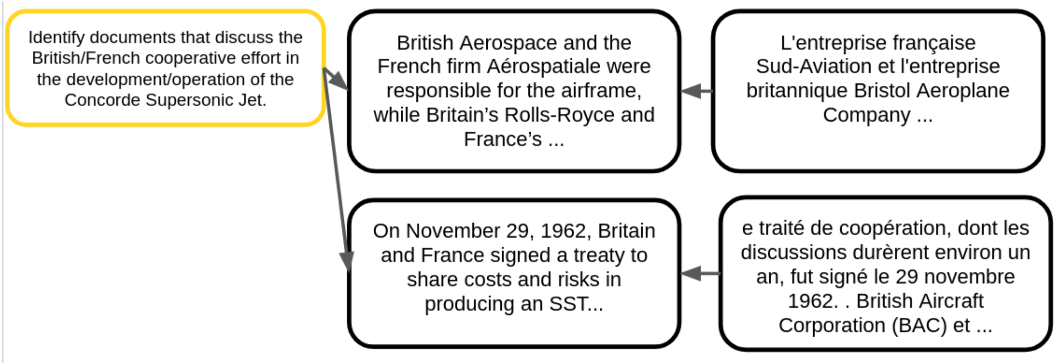
Query Translation



Query translation cont.

- Query translation is widely used in CLIR experiments, specifically because it is efficient
- But efficiency considerations may come out differently when the query workload is very high, as is the case in Web search
- Query translation also has advantages for applications in which there are many possible query languages, but only one document language
- Word sequences in queries are often very short, and thus possibly less informative

Document translation

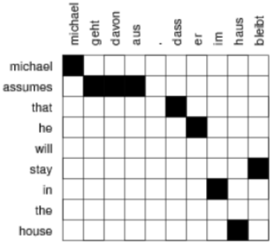


Document translation cont.

- Document translation is often effective as the words in a document occur in sequence, and we have good computational models for leveraging such sequences to make more accurate translations
- But translation direction might influence the translation quality (e.g. translating from a language with no given word boundaries, such as Chinese or Japanese, might be harder than translating into such language)
- Another obvious advantage to document translation is that if the translations are produced, then readers who are unable to read retrieved documents in their original language can be shown cached translations

Probabilistic Structured Queries (PSQ)

- PSQ uses translation matrix



English to German

P (das Haus | house) = 0.5
P (der Haushalt | house) = 0.3
P (das Parlament | house) = 0.2

$$tf(e, d_k) = \sum_{f_i} p(f_i|e) \times tf(f_i, d_k)$$

Probabilistic Structured Queries (PSQ)

- We are interested in the probability that a query term is a translation of a document term
- We can then ask, for every document, what would be the expected counts of the query terms if the document had been written in the query language.
- The probability that document term f_i (e.g., das Haus) would translate to query term e (e.g., house) might be estimated in many ways, but by far the most popular approach when parallel (i.e., translation-equivalent) text is available has been to perform term-level alignment and then compute the maximum likelihood estimate for the probabilities

Other translation possibilities

- We can translate both the queries and the documents into a common pivot language
- For example, if we want to use Navajo queries to search Burmese content we might translate both the Navajo and Burmese into English, simply because there are more language resources for Navajo-English and for Burmese-English than there are for Navajo-Burmese
- Bilingual or multilingual embeddings, can also be thought of as language-independent representations of meaning

Material project

Material: example summary: “food shortage”

Machine Translation Summary

```
CLOSE MATCH (food shortage):
...can cause stress, some food products, air change, lack of
food, misunderstandings, as well as many other factors.
Klasterinis headache pain It is quite rare, strong headache,
which is more widespread between men than women. Klasterinis
headache may arise one time during the day...
```

Human Translation Summary

```
CLOSE MATCH (food shortage):
...increases; it's often hereditary (to family members in one
or several generations); a migraine can be caused by stress,
some food products, changes in weather, lack of food, insomnia
, also numerous other factors. cluster headache it's a quite
rare, severe headache that's more prevalent among...
```

Material: Language specifics

- Methods such as Byte-Pair Encoding and Wordpiece, for example, diminished the need for language specific tokenization
- Practical systems, by contrast, can still benefit substantially from language-specific processing
- For example, in English, stemming is normally applied only to suffixes, in Arabic, by contrast, stemming must pay attention to both prefixes and suffixes, and the agglutination means that Finnish stemming will benefit from recursion

Material: CLIR character normalization

	Somali/Swahili/Tagalog
a	a
á	a
à	a
ă	a
A	a
Á	a
À	a
Ă	a
b	b
B	b
c	c
C	c
d	d
D	d

Material: CLIR character normalization cont.

- Choosing one form for letters of multiple forms

ك → ک　　گ → ګ　　ئ → ئ　　۶ → ء

ھ → ه　　ـہ → ه　　۴ → ٤　　۵ → ٥

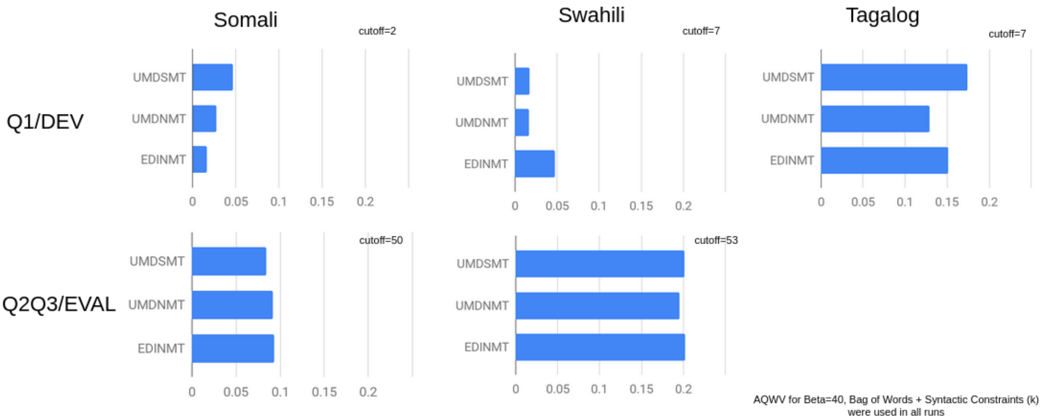
- Choosing one representative letter for letters that are used interchangeably

ى, ى, ى, ے, ې → ي　　أ, إ, آ → ا　　ة → ه

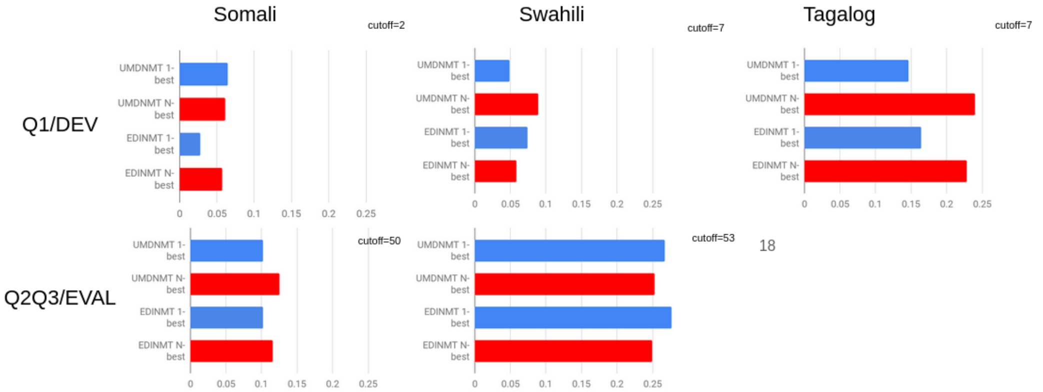
- Converting borrowed letters into their Pashto cognets

ق → ک　　ف → پ

Material: Document translation comparison



Material: N-best translation

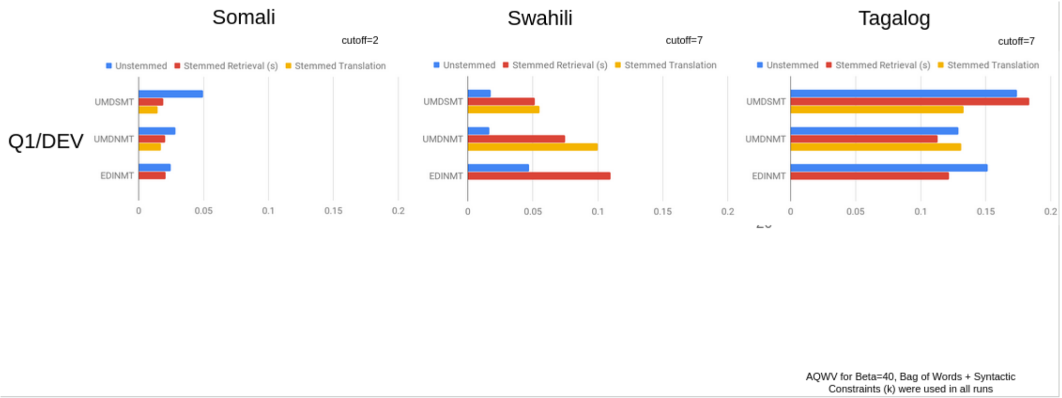


AQWV for Beta=40, Bag of Words (b) were used in all runs

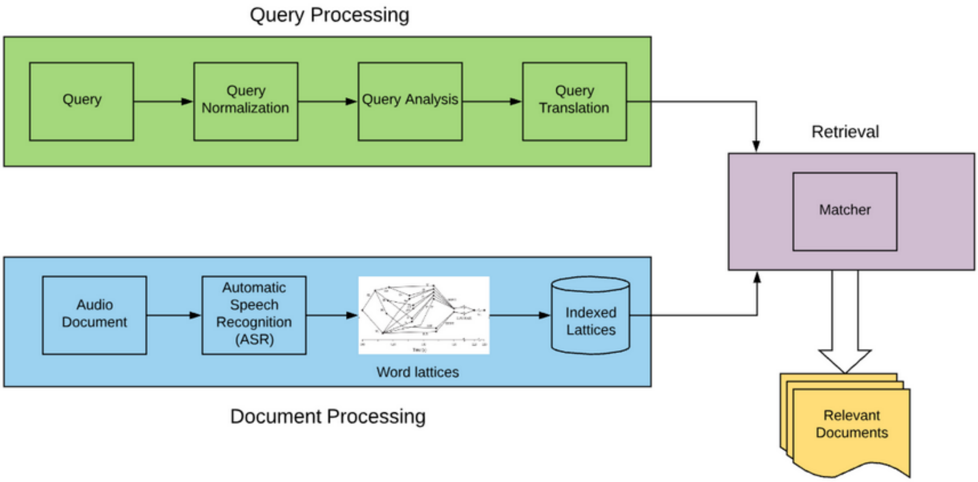
Material: Query translation results

AQWV for Beta=40, Bag of Words (b) were used in all runs

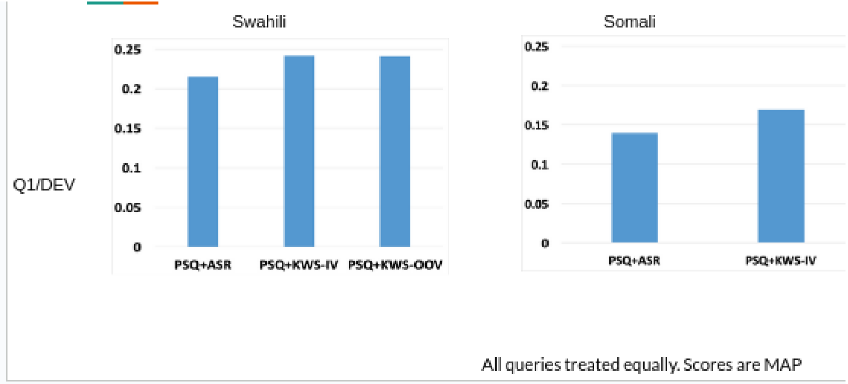
Material: Stemming



Material: Speech processing



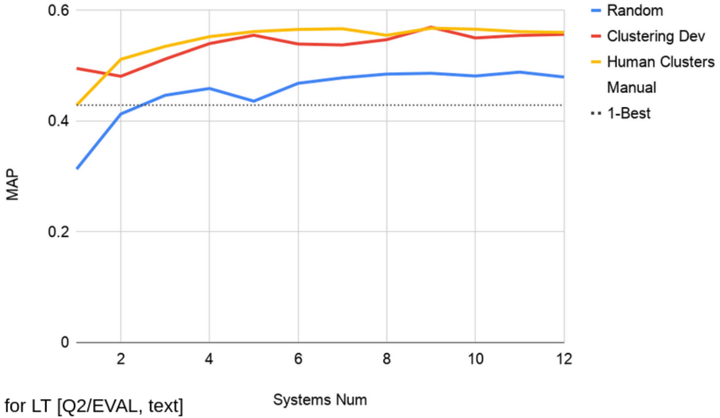
Material: Speech processing



Material: System combination

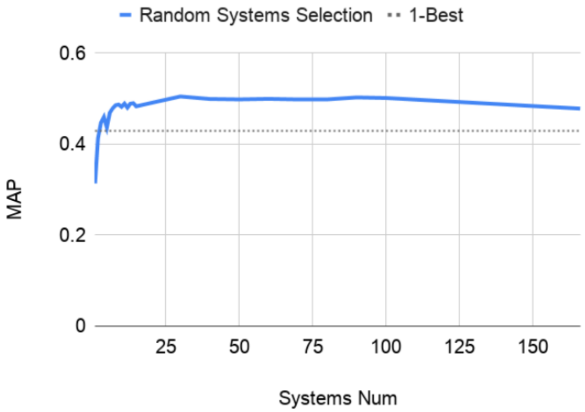
- CLIR systems have even greater potential for benefiting from diversity in a fusion than do monolingual IR systems because of the additional potential for diversity that translation resources, and ways of using those translation resources, introduces
- Late fusion combination between query and document translation, yields the best results

Material: System combination approaches



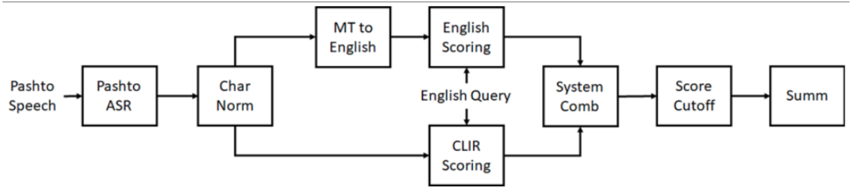
MAP scores for LT [Q2/EVAL, text]
10,203 documents, 1,000 queries

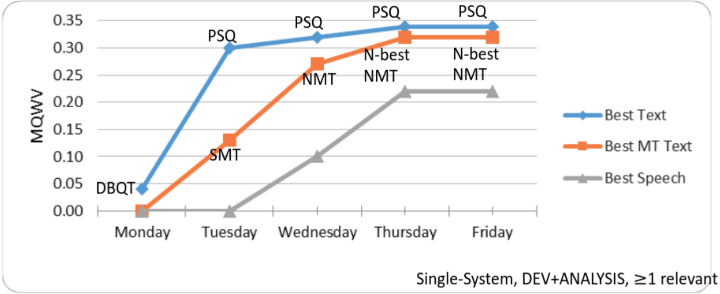
Material: System combination - random selection



MAP scores for LT [Q2/EVAL, text], WeightCombMNZ with STO normalization, 10,203 documents, 1,000 queries

Material: 5 days exercise





Summary

- MT Approaches (Rule-based, SMT, NMT)
- CLIR Approaches
- Real world CLIR example

Reading and References

- *MT Talks*[7]
- *Statistical Machine Translation*, Ondrej Bojar[8]
- *Neural Machine Translation*, Philipp Koehn[9]

[7] https://mttalks.ufal.ms.mff.cuni.cz/index.php/MT_Talks
[8] https://ufal.mff.cuni.cz/courses/npfl087#lectures
[9] http://mt-class.org/jhu/assets/nmt-book.pdf