



Embracing Diversity: Searching over multiple languages

Tommaso Teofili
Suneel Marthi

June 12, 2017
Berlin Buzzwords, Berlin, Germany

\$WhoAreWe

Tommaso Teofili

 @tteofili

- Software Engineer, Adobe Systems
- Member of Apache Software Foundation,
- PMC Chair, Apache Lucene
- Committer and PMC on Apache Joshua, Apache OpenNLP, Apache JackRabbit

Suneel Marthi

 @suneelmarthi

- Principal Software Engineer, Office of Technology, Red Hat
- Member of Apache Software Foundation
- Committer and PMC on Apache Mahout, Apache OpenNLP, Apache Streams

Agenda

- What is Multi-Lingual Search ?
- Why Multi-Lingual Search ?
- What is Statistical Machine Translation ?
- Overview of Apache Joshua
- Dataflow Pipeline
- Demo

What is Multi-Lingual Search ?



- Searching
 - over content written in different languages
 - with users speaking different languages
 - both
- Parallel corpora
- Translating queries
- Translating documents

Why Multi-Lingual Search ?

Embracing diversity

- Most online tech content is in English
 - Wikipedia dumps:
 - en: 62GB
 - de: 17GB
 - it: 10GB
- Good number of non-English speaking users
- A lot of search queries are composed in English
- Preferable to retrieve search results in native language
- ... or even to consolidate all results in one language


UC1 – tech domain, native first

Google  

[Tutti](#) [Immagini](#) [Video](#) [Shopping](#) [Notizie](#) [Altro](#) [Impostazioni](#) [Strumenti](#)

Circa 2.890.000 risultati (0,67 secondi)

(EN) A.B. Ivanov, Inner **product**, in Encyclopaedia of Mathematics, Springer e European Mathematical Society, 2002. (EN) Explanation of **dot product** including with complex vectors, mathreference.com.



[betterexplained.com](#)

[Prodotto scalare - Wikipedia](#)
https://it.wikipedia.org/wiki/Prodotto_scalare

Informazioni su questo risultato • Feedback

[Prodotto scalare - Wikipedia](#)
https://it.wikipedia.org/wiki/Prodotto_scalare ▼
In matematica, in particolare nel calcolo vettoriale, il prodotto scalare è un'operazione binaria EN)
Jeffreys, H. and Jeffreys, B. S. "**Scalar Product**." §2.06 in ...
[Definizione](#) · [Prodotto scalare nello spazio ...](#) · [Espressione analitica](#) · [Applicazioni](#)

[Dot product - Wikipedia](#)
https://en.wikipedia.org/wiki/Dot_product ▼ [Traduci questa pagina](#)
In mathematics, the **dot product** or **scalar product** is an algebraic operation that takes two equal-

UC2 – native only ?

The image shows a Google search interface with the query "host of angels". A filter menu is open, showing options like "All results", "Sites with images", "Visited pages", "Not yet visited", "Dictionary", "Reading level", "Personal", "Nearby", "Translated foreign pages", and "Verbatim". A mouse cursor is pointing at "Translated foreign pages".

The search results include a snippet for "Heavenly host" from a Bible website, with the text: "2:13 A multitude of the heavenly host; i.e. angels, who are represented as a host surrounding the throne of God (1Ki 22:19 2Ch 18:18 Ps 103:21 Da 7:10 Mt ...".

The second search result is from the IARPA website, titled "Machine Translation for English Retrieval of Information in Any Language (MATERIAL)". The page content includes:

Office of the Director of National Intelligence
IARPA
BE THE FUTURE

RESEARCH PROGRAMS | OUR PROGRAM MANAGERS | WORKING WITH IARPA | CAREERS | NEWSROOM | ABOUT IARPA

Home > Research Programs > MATERIAL > MATERIAL BAA

Machine Translation for English Retrieval of Information in Any Language (MATERIAL)

The MATERIAL performers will develop an "English-in, English-out" information retrieval system that, given a domain-sensitive English query, will retrieve relevant data from a large multilingual repository and display the retrieved information in English as query-biased summaries. MATERIAL queries will consist of two parts: a domain specification and an English word (or string of words) that capture the information need of an English-speaking user, e.g., "zika virus" in the domain of GOVERNMENT vs. "zika virus" in the domain of HEALTH, or "asperger's syndrome" in the domain of EDUCATION vs. "asperger's syndrome" in the domain of SCIENCE. The English summaries produced by the system should convey the relevance of the retrieved information to the domain-limited query to enable an English-speaking user to determine whether the document meets the information needs of the

Solicitation Status: OPEN
IARPA-BAA-16-11
Proposers' Day Date: September 27, 2016
BAA Release Date: January 19, 2017
BAA Question Period: January 19, 2017 - February 20, 2017
Proposal Due Date: March 20, 2017

Additional Information
Project Description: ...

What is Machine Translation ?

Generate Translations from Statistical Models trained on Bilingual Corpora.

Translation happens per a probability distribution

$p(e/f)$

E = string in the target language (English)

F = string in the source language (Spanish)

$e\sim = \operatorname{argmax} p(e/f) = \operatorname{argmax} p(f/e) * p(e)$

$e\sim$ = best translation, the one with highest probability

Word-based Translation

How to translate a word → lookup in dictionary
Gebäude – building, house, tower.

Multiple translations

some more frequent than others

for instance: house and building most common

Look at a parallel corpus
(German text along with English translation)

| Translation of Gebäude | Count | Probability |
|-------------------------------|--------------|--------------------|
| house | 5.28 billion | 0.51 |
| building | 4.16 billion | 0.402 |
| tower | 9.28 billion | 0.09 |

Alignment

- In a parallel text (or when we translate), we align words in one language with the word in the other

| | | | |
|-----|----------|-----|------|
| Das | Gebäude | ist | hoch |
| ↓ | ↓ | ↓ | ↓ |
| the | building | is | high |

- Word positions are numbered 1–4

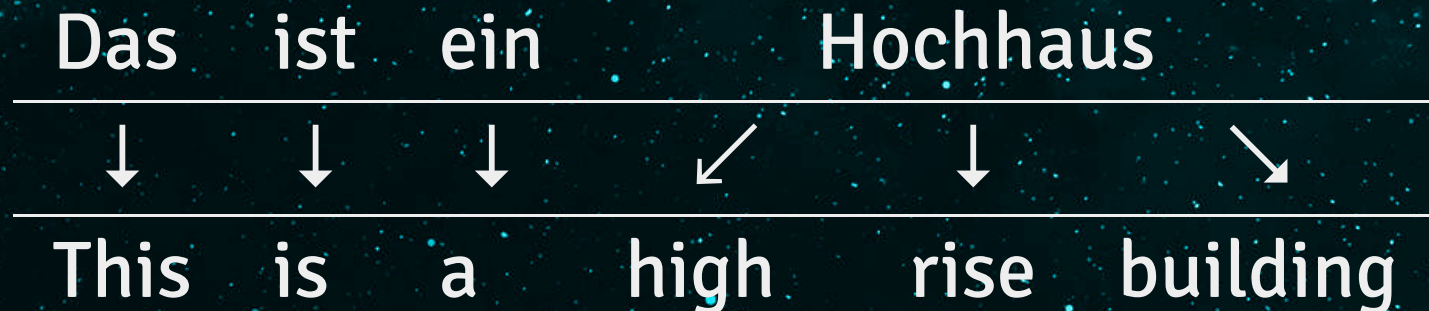
Alignment Function

- Define the Alignment with an Alignment Function
- Mapping an English target word at position i to a German source word at position j with a function $a : i \rightarrow j$
- Example

$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4\}$

One-to-Many Translation

A source word could translate into multiple target words



Phrase-based Translation

Alignment Function

- Word-Based Models translate words as atomic units
- Phrase-Based Models translate phrases as atomic units
- Advantages:
 - many-to-many translation can handle non-compositional phrases
 - use of local context in translation
 - the more data, the longer phrases can be learned
- “Standard Model”, used by Google Translate and others

Phrase-Based Model

| | | | | | |
|--------|-----|-----|----------------|--------------------------|---|
| Berlin | ist | ein | herausragendes | Kunst- und Kulturzentrum | . |
| ↓ | ↓ | ↓ | ↓ | ↓ | ↓ |
| Berlin | is | an | outstanding | Art and cultural center | . |

- Foreign input is segmented in phrases
- Each phrase is translated into English
- Phrases are reordered

Decoding

- We have a mathematical model for translation $p(e | f)$
- Task of decoding: find the translation e_{best} with highest probability

$$e_{\text{best}} = \operatorname{argmax} p(e | f)$$

- Two types of error
 - the most probable translation is bad → fix the model
 - search does not find the most probable translation → fix the search

Translation Process

Translate this query from German into English

er trinkt ja noch nichts

er



he

Pick and input phrase, translate

Translation Process

Translate this query from German into English

er trinkt ja noch nichts

er

ja noch nichts



he



does not yet

Pick and input phrase, translate

Translation Process

Translate this query from German into English

er trinkt ja noch nichts

er

trinkt

ja noch nichts



he




does not yet

drink

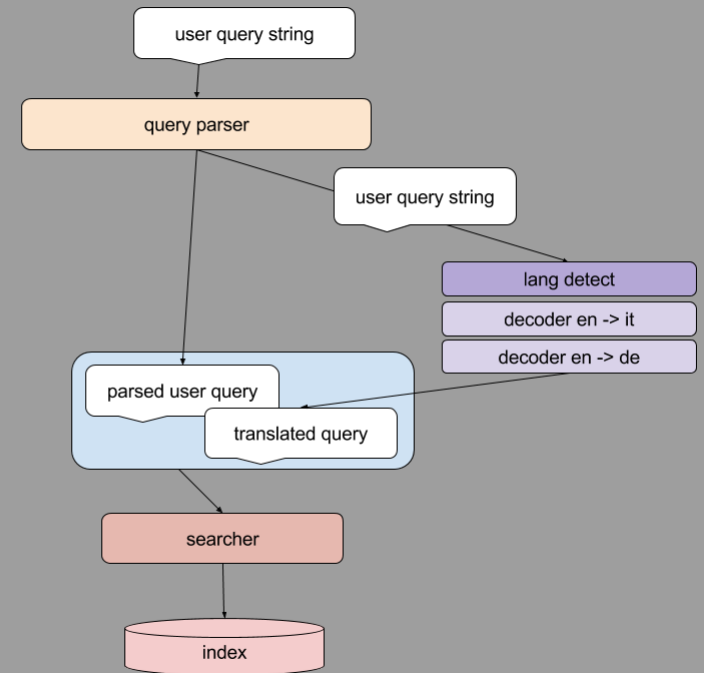
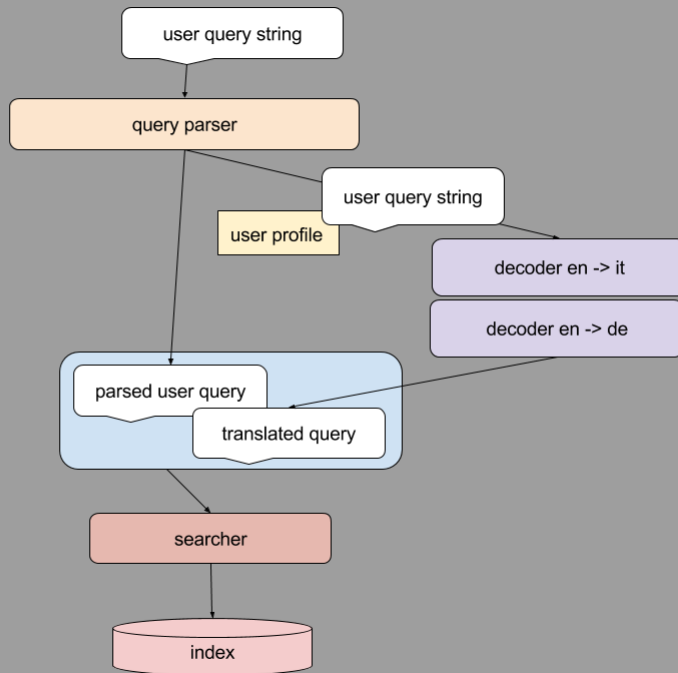
Pick and input phrase, translate

Apache Joshua



- Statistical Machine Translation Decoder for phrase-based and hierarchical machine translation
- Written in Java
- Provide 64 language packs for machine translation
 - [https://cwiki.apache.org/confluence/display/JOSHUA/Language+P](https://cwiki.apache.org/confluence/display/JOSHUA/Language+Packs)
- Project initiated by Johns Hopkins Univ. and University of Pennsylvania
- Presently incubating at Apache Software Foundation
- Used extensively by Amazon.com, NASA JPL
- <https://cwiki.apache.org/confluence/display/JOSHUA>
-  @ApacheJoshua

Flows



References

- Apache Joshua — <https://cwiki.apache.org/confluence/display/JOSHUA>
- Apache OpenNLP — <https://opennlp.apache.org>
- GitHub — <https://github.com/smarthi/BBuzz-multilang-search>
- Slides — <https://smarthi.github.io/bbuzz17-embracing-diversity-searching-over-multiple-languages/#/>

Credits

- Joern Kottmann – PMC Chair, Apache OpenNLP
- Matt Post – PMC Chair, Apache Joshua
- Bruno P. Kinoshita – Committer on Apache OpenNLP, committer and PMC on Apache Commons and Apache Jena

Questions ???