Advanced Methods of Information Retrieval
- Multilingual Data Analytics and Retrieval -

Simon Gottschalk
SS 2018
April 26, 2018
Recap:
Information Extraction
Named Entity Extraction

- **Named Entities**: Proper nouns or phrases, which refer to real-world objects (entities).

- **Named Entity Extraction (Recognition, Identification)**: Detecting boundaries of named entities (NEs) in unstructured text:

  Dubrovnik is a Croatian city on the Adriatic Sea, in the region of Dalmatia founded in the 7th century. The Imperial Fortress was built in 1806 by Marshal Marmont in honor of emperor Napoleon. The HBO series Game of Thrones used

- **Named Entity Classification**: Automatically assigning pre-defined classes to NEs, such as PERSON, LOCATION, ORGANISATION:

  LOCATION  PERSON
  ORGANIZATION  PERCENT
  DATE  TIME
  MONEY
Entity Linking

- **Entity Linking (EL):** detecting entities and linking them to the entries of a Knowledge Base

*Dubrovnik* is located in the region of Dalmatia.

→ http://dbpedia.org/resource/Dubrovnik

→ https://www.wikidata.org/wiki/Q1722
Temporal Extraction

- **Temporal extraction** is the extraction and normalization of temporal expressions.

After World War I, Dubrovnik became part of Croatia which itself was part of the Kingdom of Serbs, Croats and Slovenes which became Yugoslavia after World War II. Dubrovnik was subjected to considerable shelling by Serbs during the war in 1991/2 in a siege that lasted seven months. The Old Town suffered considered damage, but was quickly restored to its former beauty.

[tagged using Heideltime]

After World War I, Dubrovnik became part of Croatia which itself was part of the Kingdom of Serbs, Croats and Slovenes which became Yugoslavia after World War II. Dubrovnik was subjected to considerable shelling by Serbs during the war in 1991/2 in a siege that lasted seven months. The Old Town suffered considered damage, but was quickly restored to its former beauty.

[tagged using SUTime]
Relation Extraction

- **Relations** between two or more entities, which relate to one another in real life. Examples:
  - located-in(Dubrovnik, Croatia)
  - married-to(Angelina Jolie, Brad Pitt).
- **Relation extraction:**
  - is a task of detecting relations between entities and assigning relation types to them.

located-in(Dubrovnik, the region of Dalmatia)
Open Information Extraction (Open IE)

- **Open IE** extracts tuples consisting of argument phrases and a relation between the arguments.

  - For example: “*Trump was elected President.*”
    - *(Trump; was elected; President)*

- Different to relation extraction
  - No pre-specified sets of relations
  - No domain-specific knowledge engineering
Multilingual Data Analytics and Retrieval
Aims of the Session

- Provide an overview of language use in the world and the Web

- Discuss which information multilingual data on the Web can provide regarding:
  - neutrality across the languages
  - cultural differences

- Provide an overview of multilingual knowledge resources
  - Wikipedia, DBpedia, Wikidata, WordNet, BabelNet

- Provide an overview of methods to compare and retrieve text across languages
Motivation I: News in Different Languages

Global Opinions

Trump has given Putin the best gift he could ask for

By Michael McFaul | May 17

Russian President Vladimir Putin speaks at a news conference in January in Moscow. (Sergei Ilnitsky/Pool Photo via Associated Press)

Certainly, the Russian government has a list of things it wants from the United States in a perfect world: lifting sanctions on Russian companies and individuals, recognition of Syrian President Bashar al-Assad as an ally in the fight against terrorism in Syria, agreement to end all dimensions of democracy and human rights, acquiescence to spheres of influence for our two countries, the granddaddy ask of them all, American recognition of Crimean “reunification.”

Judging by the jolly photos published by Russian media of Russian Foreign Minister Lavrov’s Oval Office meeting with President Trump, it didn’t look as though Trump was discussing the tough topics: Russia’s invasion of Ukraine, Crimean annexation, Russian opposition leaders Alexei Navalny and Vladimir Kara-Murza, Aleppo’s obliter, Russia’s violation of our sovereignty during last year’s presidential election. In his

Putin ist für Trump der ungünstigste Entlastungszeuge

Von Ansgar Graw, Julija Smirnova, Washington, Moskau | Stand: 18.05.2017 | Lesedauer: 8 Minuten

In der Affäre um angeblichen Geheimnissverrat bietet Putin Trump seine Hilfe an. Das dürfte in der US-Besiedlung ausgesprochen schlecht ankommen. Zunehmend gehen auch Republikaner auf Distanz zum Präsidenten,
Motivation II: European Food Culture on Wikipedia

[Graph showing the similarity of cuisines across different languages.]

- Bulgarian
- Bosnian cuisine
- Catalan cuisine
- Czech cuisine
- Danish cuisine
- Austrian cuisine
- German cuisine
- British cuisine
- English cuisine
- Irish cuisine
- Spanish cuisine
- Estonian cuisine
- Finnish cuisine
- French cuisine
- Croatian cuisine
- Hungarian cuisine
- Italian cuisine
- Lithuanian cuisine
- Latvian cuisine
- Belgian cuisine
- Dutch cuisine
- Norwegian cuisine
- Polish cuisine
- Portuguese cuisine
- Romanian cuisine
- Russian cuisine
- Slovak cuisine
- Serbian cuisine
- Swedish cuisine
- Turkish cuisine
- Ukrainian cuisine

Legend:
- Austrian cuisine
- Belgian cuisine
- Bosnian cuisine
- British cuisine
- Bulgarian cuisine
- Catalan cuisine
- Croatian cuisine
- Czech cuisine
- Danish cuisine
- Dutch cuisine
- English cuisine
- Estonian cuisine
- Finnish cuisine
- French cuisine
- German cuisine
- Hungarian cuisine
- Irish cuisine
- Italian cuisine
- Latvian cuisine
- Lithuanian cuisine
- Norwegian cuisine
- Polish cuisine
- Portuguese cuisine
- Romanian cuisine
- Russian cuisine
- Slovak cuisine
- Serbian cuisine
- Swedish cuisine
- Turkish cuisine
- Ukrainian cuisine

Catalonia understands Iberian cuisines.
France understands Scandinavia.
George Washington was an American statesman and soldier who served as the first President of the United States from 1789 to 1797 and was one of the Founding Fathers of the United States.

乔治·华盛顿，美国国父，1775年至1783年美国独立战争时的殖民地军总司令，1789年成为美国第一任总统（其同时也成为全世界第一位以“总统”为称号的国家元首），在接连两次选举中都获得全票选票团无异议支持，一直担任总统直到1797年。他也是一名共济会成员。
**Motivation IV: Cross-lingual Text Similarity**

<table>
<thead>
<tr>
<th>George William Gray</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Education and career</strong></td>
</tr>
<tr>
<td>Born in Denny, Scotland, Gray was educated at the University of Glasgow and while working as an assistant lecturer at the University College in Hull (then part of the University of London) obtained his PhD in 1953. He developed his academic career at the college, which became the University of Hull in 1954, from 1946 to 1990. He was appointed senior lecturer in 1960, Professor of Organic Chemistry in 1974, and GF Grant Professor of Chemistry in 1984. He remained an Emeritus Professor at Hull. In 1990 he joined the chemical company Mercis, then became an independent consultant in 1996.</td>
</tr>
<tr>
<td><strong>Liquid crystals</strong></td>
</tr>
<tr>
<td>In 1973, in conjunction with the Royal Radar Establishment, he showed that 4-Cyano-4-n-pentylbiphenyl possessed a stable nematic phase at room temperature. This compound and other long-lasting cyanobiphenyls made the twisted nematic display (LCD) popular. Gray</td>
</tr>
</tbody>
</table>

---

Language Use in the World and in the Web
Languages in the World

- Chinese is the most frequent native language, but English is studied by most people

Of the 7.2 billion people on Earth...

- Arabic 467M
- Spanish 389M
- Chinese (all dialects) 1.39 billion speakers
- Hindi-Urdu 588M
- Bengali 250M
- Portuguese 193M
- Italian 67M
- German 132M
- Japanese 123M
- French 118M
- Russian 254M

English is by far the most common studied foreign language in the world.

French 82M

English 1.5 billion learners

Chinese 30M

Spanish 14.5M

German 14.5M

Italian 8M

Japanese 3M

Sources: Ulrich Ammon, University of Düsseldorf
Note: Totals for languages include bilingual speakers.

THE WASHINGTON POST

Usage of Content Languages for Websites

https://w3techs.com/technologies/history_overview/content_language
Number of Internet Users by Language

- English
- Chinese
- Spanish
- Arabic
- Portuguese
- Indonesian
- French
- Japanese
- Russian
- German
- Others

% Internet Users of World Total

Languages in the Web

- big discrepancies between the number of users and websites:
  - although the number of English-speaking users is similar to those of
    Chinese speaking users, less than 2% of the websites have
    Chinese content (English: 52.1%)

- “only” 52.1% of the websites are English
  - need for analytics of web content in different languages
  - long tail of rare languages that need support for information
    extraction, retrieval, …

- temporal evolution shows that the percentage of non-English is
  increasing faster
  - the number of Chinese Web users has increased by 2390% since
    2000 (English: 648%)
Terminology

- **Multilingual Information Access (MLIA):**
  - MLIA addresses the problem of accessing, querying and retrieving information from collections in any language at any level of specificity.

- **Cross-lingual/Translingual Information Retrieval (CLIR/TLIR):**
  - Cross–language information retrieval (CLIR) is a subfield of the traditional information retrieval (IR). It provides users with access to information that is in a different language from their queries.
ISO Language and Country Codes

- ISO 639 series of International Standards
  - ISO 639-1: 2 letter code for most of the major languages
  - ISO 639-2: 3 letter code, which covers more languages
  - ISO 639-3: 3 letter code that aims to provide all languages, including macro, living, extinct and ancient languages

<table>
<thead>
<tr>
<th>Language Name</th>
<th>ISO 639-1</th>
<th>ISO 639-2</th>
<th>ISO 639-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>German (Deutsch)</td>
<td>de</td>
<td>deu</td>
<td>deu</td>
</tr>
<tr>
<td>English</td>
<td>en</td>
<td>eng</td>
<td>eng</td>
</tr>
<tr>
<td>Balinese</td>
<td>-</td>
<td>ban</td>
<td>ban</td>
</tr>
<tr>
<td>Babatana</td>
<td>-</td>
<td>-</td>
<td>baa</td>
</tr>
<tr>
<td>Mandarin Chinese (language group/macro language)</td>
<td>-</td>
<td>-</td>
<td>cmn</td>
</tr>
</tbody>
</table>

- ISO 3166: Country Codes
  - ISO 3166-1 alpha-2: two-letter country codes, used as top-level domains
    - DE: Germany, GB: United Kingdom, US: United States of America

https://www.iso.org/iso-639-language-codes.html
Languages on the Web

- **IETF language tags**
  - language codes defined by the Internet Engineering Task Force (IETF)
    - British English: en-GB, American English: en-US

- **i18n: internationalization:**
  - designing a software application so that it can be adapted to various languages and regions without engineering changes

- **L10n: Localization**
  - adapting internationalized software for a specific region or language by adding locale-specific components and translating text

- **IRI: Internationalized Resource Identifier**
  - URIs are restricted to ASCII characters, IRIs allow Unicode
    - https://en.wiktionary.org/wiki/%E1%BF%AC%CF%8C%CE%B4%CE%BF%CF%82
    - vs https://en.wiktionary.org/wiki/Πόδος
The Multilingual Wikipedia
Wikipedia Language Editions

- there are approximately 300 Wikipedia language editions
- each language in Wikipedia has a separate set of user accounts
- URLs based on ISO 639-1 and ISO 639-3 (e.g. de.wikipedia.org)

Distribution of articles in language editions

<table>
<thead>
<tr>
<th>Language</th>
<th>#Articles</th>
<th>#Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>5,620,416</td>
<td>33,431,934</td>
</tr>
<tr>
<td>German</td>
<td>2,174,918</td>
<td>2,889,767</td>
</tr>
<tr>
<td>Greek</td>
<td>145,967</td>
<td>243,978</td>
</tr>
</tbody>
</table>

(the majority of the articles in Swedish, Cebuano, and Waray were created by the bot “Lsjbot”)

Interlingual Links

- Wikipedia pages in different language editions are interlinked using interlingual links
- since 2013, these links are maintained in Wikidata
Concept Overlap across Languages

- as of 2010, over 74 percent of concepts were described in only one language
- more than 95.5 percent of concepts appeared in six or fewer languages

Hecht, Brent, and Darren Gergle. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. SIGCHI ’10.
Neutrality in Wikipedia - NPOV

Wikipedia: “All encyclopedic content on Wikipedia must be written from a **neutral point of view (NPOV)**, which means representing fairly, proportionately, and, as far as possible, **without editorial bias**, all of the significant views that have been published by reliable sources on a topic.

**NPOV is a fundamental principle** of Wikipedia and of other Wikimedia projects. It is also one of Wikipedia's three core content policies…”

A Linguistic Point of View (LPOV) is a language-specific representation of a topic.

LPOV can become apparent when different communities develop their own divergent representations for the same topic.

LPOVs can be observed in Wikipedia, given the relative isolation of communities of editors of each Wikipedia language edition.

Using Wikipedia to Understand Cultures

- Wikipedia as a tool to conduct cultural research
- comparative analysis of articles on the same subject matter across language versions
- particularly interesting in the case of controversial articles
  - Politics, Religion, …

- example case study: Massacre of Srebrenica that affected several different language communities

Feature: Images

- images are language-independent and therefore easier to compare

- the selection of images can give an idea about the relevance of topics in a language context

- Srebrenica example: Bosnian article has images about the event itself, others emphasize the investigations

Feature: Editor Locations

- investigating the editors’ locations can help to understand how much *cross-editing* is involved between the languages

- editors tend to be dedicated to one language, with the exception of the English articles

Feature: Temporal Evolution

- the whole evolution of a Wikipedia article can be reconstructed
- title changes according to editor discussions
  - Drama of Srebrenica, Srebrenica genocide, Srebrenica massacre, …
- example of Brexit: timeline of edits, and article creation dates

More Wikipedia Features

- tables of contents
  - approximation of the content
- discussion pages
  - controversies between editors become evident on the discussion pages
- references
  - citing the same news \(\rightarrow\) same point of view?
- links
  - compute overlap of links (via interlingual links)
- ...
Multilingual Knowledge Graphs
Cross-lingual Information in Knowledge Graphs

<table>
<thead>
<tr>
<th>DBpedia</th>
<th>Freebase</th>
<th>Open-Cyc</th>
<th>Wikidata</th>
<th>YAGO</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Main” DBpedia is English (properties etc.), but linked localized versions are available in 125 languages (localized are textual descriptions such as rdfs:label, rdfs:comment, dbpedia-owl:abstract. There are also links to local versions of Wikipedia)</td>
<td>Human readable IDs are in English, but every entity and property has an i18n in many languages</td>
<td>English</td>
<td>Almost every language (by community), including dialects</td>
<td>All entity names are from English Wikipedia, some rdfs:label values have different languages</td>
</tr>
</tbody>
</table>

**Wikidata**

- one KG for all languages
- resource and property identifiers are language-independent
- labels, aliases and descriptions available in different languages
- each Wikipedia page in any language is mapped to a Wikidata item
DBpedia Language Editions

- one language-specific DBpedia is extracted from each Wikipedia language edition
- for some languages, there are canonically aligned DBpedia editions using Wikidata IDs:
  - localized: <http://de.dbpedia.org/resource/Ang_Lee>
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
    <http://dbpedia.org/ontology/Person> .
  - canonicalized: <http://wikidata.dbpedia.org/resource/Q160726>
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
    <http://dbpedia.org/ontology/Person> .

- some properties are canonicalized:

- labels are provided in several languages
- no integration of relations between languages
Cross-lingual SPARQL Queries

- each literal can be combined with an optional language tag

```
SELECT ?label WHERE {
}
```
Cross-lingual SPARQL Queries

- Language tags can be queried directly or with the string and language part:
- Same result (http://dbpedia.org/resource/Hanover) for:

```
SELECT ?city WHERE {
  ?city rdfs:label "Hannover"@de.
}
```

```
SELECT ?city WHERE {
  FILTER(STR(?label) = "Hannover") .
  FILTER(LANGMATCHES(LANG(?label), "de")) .
}
```
WordNet

- large lexical database of English
- nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept

- WordNets are available in other languages, mainly maintained by universities
- list of Wordnets is collected by the Global WordNet Association

http://www.globalwordnet.org/gwa/wordnet_table.html

English WordNet for “fish”
Wiktionary

- Wikipedia sister project run by the Wikimedia Foundation, written collaboratively by volunteers
- available in 171 languages
- each Wiktionary page represents a word
- definitions, synonyms, declension, …
- mapping between languages
  - language links as in Wikipedia
  - translations for each word sense

<table>
<thead>
<tr>
<th>Person</th>
<th>Wortform</th>
</tr>
</thead>
<tbody>
<tr>
<td>ich</td>
<td>fische</td>
</tr>
<tr>
<td>du</td>
<td>fischst</td>
</tr>
<tr>
<td>er, sie, es</td>
<td>fischt</td>
</tr>
</tbody>
</table>

| Präteritum | ich | fischte |
| Konjunktiv II | ich | fischte |

<table>
<thead>
<tr>
<th>Imperativ</th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fische!</td>
<td>fischt!</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Perfekt</th>
<th>Partizip II</th>
<th>Hilfsverb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>gefischt</td>
<td>haben</td>
</tr>
</tbody>
</table>

German Wiktionary (“fischen”)

- Dänisch: [1] fiske → da
- Englisch: [1] fish → en
- Färöisch: [1] fiska → fo

English Wiktionary (“fish”)

**Verb** [edit]

fish (third-person singular simple present fishes, present participle fishing, simple past and past participle fished)

1. (intransitive) To hunt fish or other aquatic animals. [quotations▼]
   
   She went to the river to fish for trout.

2. (transitive) To search (a body of water) for something other than fish.

   They fished the surrounding lakes for the dead body.

3. (intransitive) To (attempt to) find or get hold of an object by searching.

**Noun** [edit]

fish (plural fishes)

1. A period of time spent fishing.

   The fish at the lake didn't prove successful.
BabelNet 4.0

- “The largest multilingual encyclopedic dictionary and semantic network”
  - 284 languages
  - 9,671,518 entities
  - 832,469,391 babel senses
- resources used to build BabelNet:
  - Wordnets in all languages
  - Wikipedia
  - Wikidata
  - Wiktionary
  - OmegaWiki (multilingual dictionary)
  - GeoNames (geographical database)
  - Wikiquote (collection of citations)
  - ...
BabelNet 4.0

- encyclopedic dictionary, semantic network and knowledge base
- each Babel synset represents a given meaning and contains all the synonyms which express that meaning in different languages

“Herring” in the BabelNet
Cross-Lingual Natural Language Processing (NLP)

- most NLP algorithms are rule-based or need to be trained on text → algorithms are language-dependent

- tokenization:
  - Dubrovnik is located in the region of Dalmatia.
  - 杜布羅夫尼克位於達爾馬提亞地區？

- many NLP tools are limited to very few languages, mostly English

<table>
<thead>
<tr>
<th>ANNOTATOR</th>
<th>AR</th>
<th>ZH</th>
<th>EN</th>
<th>FR</th>
<th>DE</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokenize / Segment</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Sentence Split</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Part of Speech</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Lemma</td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Named Entities</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Constituency Parsing</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Dependency Parsing</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mention Detection</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coreference</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Open IE</td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Stanford CoreNLP’s language support

https://stanfordnlp.github.io/CoreNLP/human-languages.html
Cross-lingual Stemming

- stemming rules are language-dependent

<table>
<thead>
<tr>
<th>Word</th>
<th>English stemming</th>
<th>German stemming</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hering (German for “herring”)</td>
<td>here</td>
<td>hering</td>
</tr>
<tr>
<td>fishing</td>
<td>fish</td>
<td>fishing</td>
</tr>
</tbody>
</table>

- stemming algorithm groups that follow similar rules:
  - Porter Stemmer: English
  - Romance (French, Spanish, Portuguese, …)
  - Germanic (German, Dutch)
  - Scandinavian (Swedish, Norwegian, Danish)
  - Russian
  - …

- implemented by the Snowball stemmer: https://snowballstem.org/
Language Identification

- Task: Given a text T, identify its language
- Training Phase: language profile creation
  - extract n-grams from a training corpus
  - rank the n-grams by descending frequency
  - example language profile for English:
    - [TH, ING, ON, ER, AND, ED]
- Identification:
  - create a document profile of T
  - example document profile:
    - [TH, ER, ON, LE, ING, AND]
  - determine the most similar language profile using rank-order statistics
  - example, with “out of place” measure:
    - \(0 + 2 + 0 + K + 3 + 1\) (K: fixed cost for an N-gram which is not found)
Monolingual and Cross-lingual Text Similarity
Monolingual Text Similarity

- Character-based
  - Longest common substring (LCS)
  - N-gram similarity
  - Levenshtein distance
- Term-based
  - Jaccard similarity
  - Dice similarity
  - Cosine similarity
- Knowledge-based
  - Similarity based on information extraction
- Corpus-based
  - Word embedding similarity
Term-based Text Similarity

- Jaccard similarity and Dice similarity
  - both texts are represented as sets of terms A and B
  - Jaccard similarity: \( J(A, B) = \frac{|A \cap B|}{|A \cup B|} \)
  - Dice similarity: \( D(A, B) = \frac{2 \cdot |A \cap B|}{|A| + |B|} \)

- Cosine similarity
  - both texts are represented as term vectors \( v \) and \( w \), based on tf-idf
  - \( \text{td-idf}: \text{term frequency–inverse document frequency} \)
  - cosine similarity: \( \text{sim}(v, w) = \frac{v \cdot w}{\|v\| \|w\|} \)

- How to apply these methods to compare two texts in different languages?
Term-based Text Similarity between Languages

- Jaccard similarity
- Dice similarity
- cosine similarity

- use a bilingual dictionary

- Problems:
  - multiple potential translations
    - “run” (en) → “laufen”, “rennen” (de)
  - translate word-by-word or find phrases
    - “tree house” (en) → “Baum Haus”, “Baumhaus” (de)
Knowledge-based Text Similarity

- semantic similarity based on the degree of similarity between words using knowledge graphs
- overlap of named entities and time expressions

In late 1753, Dinwiddie ordered Washington to deliver a letter asking the French to vacate the Ohio Valley.


Time Overlap: 100%, Entity Overlap: $\frac{2}{5} = 40\%$ (Jaccard similarity)

- use WordNet, BabelNet, …
Corpus-based Text Similarity: Word Embeddings

https://deeplearning4j.org/word2vec.html
Word Embeddings

- given: set of terms $V$ (vocabulary)

- traditional way of representing words: one-hot vector
  - size of each vector: vocabulary size $|V|$

- word embeddings: reduce dimension of the vector space
  - intuition: words that occur in similar contexts tend to have similar meanings
  - vector size is configured (often $d = 300$)

$W(\text{“woman”}) - W(\text{“man”}) \approx W(\text{“queen”}) - W(\text{“king”})$

https://www.tensorflow.org/images/linear-relationships.png
Bilingual Word Embeddings

- embed the terms in both languages in the same vector space
- similarity between two words: cosine similarity of embeddings

Luong, Thang, Hieu Pham, and Christopher D. Manning. "Bilingual word representations with monolingual quality in mind." Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. 2015.
Training Bilingual Word Embeddings

- cross-lingual word embedding models employ two different kinds of data:
  - Type of alignment: Models use different types of bilingual signals, which introduce stronger or weaker supervision
  - Comparability: Models require either parallel data sources, that is, exact translations in different languages or comparable data that is only similar in some way

<table>
<thead>
<tr>
<th>Alignment Level</th>
<th>Parallel</th>
<th>Comparable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Dictionaries</td>
<td>Images</td>
</tr>
<tr>
<td>Sentence</td>
<td>Translations</td>
<td>Captions</td>
</tr>
<tr>
<td>Document</td>
<td></td>
<td>Wikipedia</td>
</tr>
</tbody>
</table>
Training Bilingual Word Embeddings

- **cat** → **chat**
- **dog** → **chien**

(a) Word, par.  (b) Word, comp.  (c) Sentence, par.  (d) Sentence, comp.  (e) Doc., comp.

---

My question relates to something that will come up on Thursday and which I will then raise again.

Meine Frage betrifft eine Angelegenheit, die am Donnerstag zur Sprache kommen wird und auf die ich dann erneut verweisen werde.

---

(c) Parallel sentence-aligned corpus: Translations from the European Parliament in >20 languages
Cross-lingual Information Retrieval
Cross-lingual Information Retrieval (CLIR)

- CLIR provides users with access to information that is in a different language from their queries

- Types of CLIR:
  - documents monolingual, queries in any language
  - documents in multiple languages, query in any language
  - documents mixed-language, query in any language

- Two Approaches:
  - query Translation
  - document Translation
CLIR — Query Translation

- queries often not complete sentences (but MT learns from that)
- merging step to rank results from different languages needed
CLIR — Document Translation

- context available
- translation done offline
- re-translations needed
- high storage demands
CLIR — Interfaces

“Unfortunately, this feature never saw much pick up — but you can still use Chrome to translate entire pages very easily, with a built-in translation bar that helps you read content on the Web, regardless of the language.”


Summary

- In this session we provided an overview of language use in the world and on the Web
  - in particular: in Wikipedia and knowledge graphs

- We discussed the role of Wikipedia’s Neutral Point of View and Linguistic Point of View in the multilingual context

- We provided an overview of methods to compare and retrieve text across languages:
  - character-based, term-based, knowledge-based, corpus-based

- We provided an overview of methods to compare and retrieve text across languages
Thank you!

Questions, Comments?

Simon Gottschalk
L3S Research Center
Leibniz University of Hannover

e-mail: gottschalk@L3S.de
References

  - Paper: Hecht, Brent, and Darren Gergle. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. SIGCHI '10.

- Language Use in the World and in the Web
References

- **Multilingual Knowledge Graphs**
References

- **Text Similarity**

- **Cross-lingual Information Retrieval**