Lecture

Knowledge-based Systems

Part 7 – Multilingual knowledge

Dr. Mohsen Mesgar

Language Technology Lab
Universität Duisburg-Essen

Recall …

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

(Artificial) intelligence: the ability acquire knowledge and adopt that knowledge to new environment and tasks to achieve goals.

Knowledge base (KB) is the core of KB systems

Two types of KBs (based on knowledge representations)

Symbolic

Connectionist

Evaluating KBs using semantic relatedness tasks

Semantic relations between words

Any other open questions?

In this lecture, you learn about …

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- **Multilingual KBs**
- **Examples of symbolic multilingual KB**
  - BabelNet
  - Parallel corpora
- **Examples of connectionist multilingual KBs**
  - Multilingual word embeddings
  - XLM

Multilinguality

Motivation

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- There are more than 6000 languages in the world
- If we consider only the top 80% of native speakers with the most frequent language, it's still 50 individual languages
- Most AI agents have been developed for English (only)

- ideally:
  - As a user, we want to access information in our own language
  - As a researcher, we do not want to re-invent the wheel for every language
    - re-use datasets from different languages
    - transfer tools to a new language

Are languages really that different? – The Innateness Hypothesis

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Children learn any first language incredibly fast, given the poverty of stimulus (Chomsky). They are just not exposed to rich enough data to acquire every single feature of their language.

Assumption: There must be some innate knowledge about language in general. Noam Chomsky called this the Universal Grammar (UG).

Concepts thought to be part of UG:

- Tense
- Number
- word order matters

→ there are commonalities between languages we can use

A simple idea to get Multilingual KBs

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- **dictionary**

- **Problem with dictionaries:**

  Often no 1:1 mapping

  - *bank* —— *Bank?*

  - *bank* —— *Bank*
    Ufer
    coast
    shore

- The translation of a word is context dependent.
- Translation is often dependent on the word sense rather than the surface form

Dictionary

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- First thing that might come to mind: dictionary

Problem with dictionaries:

Meaning is compositional:

- *Die Daumen drücken* ..... ~~*To press the thumbs?*~~
- ~~*Die Finger gekreuzt*~~ ..... *To keep your fingers* ~~*halten?*~~ *crossed*

- A translation is more than the sum of word by word translations

Symbolic Multilingual KBs

BabelNet

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

BabelNet

Mutter | GERMAN | ENGLISH | ÜBERSETZEN

PRÄFERENZEN

Alle | Konzepte | Named Entities | 11 konzepte

- Nomen

Nomen

Mutter, Mama, Mutterschaft

Mutter bezeichnet den weiblichen Elternteil einer Person.

ID: 00034027n | Konzept

EN mother, female parent, mothers

Nonne, Ordensschwester, Mutter

Als Nonne bezeichnet man ein weibliches Mitglied mancher christlicher Ordensgemeinschaften sowie eines buddhistischen Ordens.

ID: 00058313n | Konzept

EN nun, sister, mother

Mutter (Technik), Schraubenmutter, DIN 934

Die Mutter, zur Abgrenzung manchmal auch Schraubenmutter, ist das mit einem Innengewinde versehene Gegenstück einer Schraube oder eines Gewindebolzens.

ID: 00058343n | Konzept

EN nut

BabelNet

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Multilingual **semantic network**
- **Integrates** knowledge from **WordNet** and **Wikipedia**

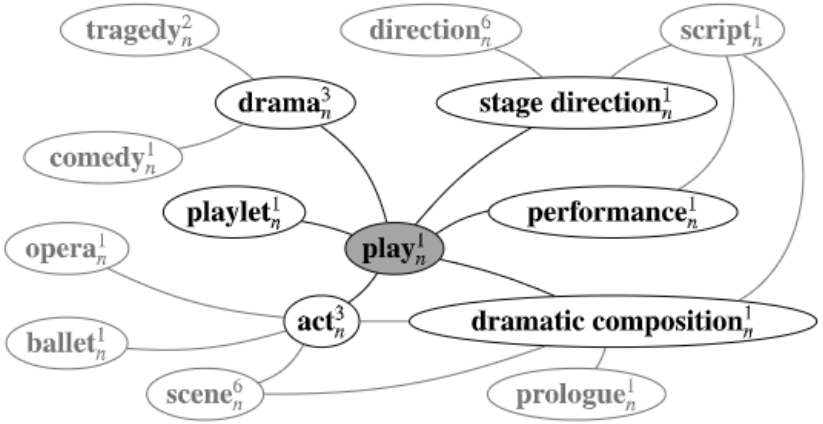
Wikipedia

UNIVERSITÄT DUISBURG ESSEN

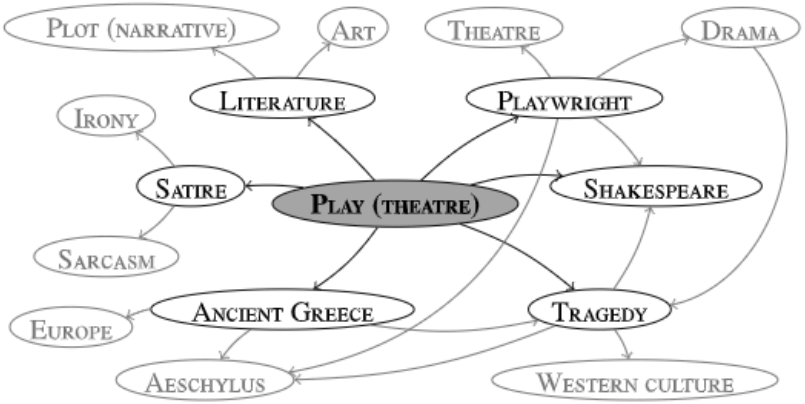
Offen im Denken

- **Wikipedia** is **structured knowledge:**
  - **Redirect Pages: Synonymy**
    Stageplay and *Theatrical play* both redirect to *Play (theatre)*.
  - **Disambiguation pages: homonymy and polysemy**
    Play links to both Play (theatre) and Play (activity)
  - **Internal links**: **related concepts**.
    Play(theatre) links to Literature, Playwright, Dialogue
  - **Inter-language links**: counterparts in other languages
    Play (theatre) links to the German Bühnenwerk.
  - **Categories**: Play(theatre) is categorized under THEATRE, DRAMA, LITERATURE, etc.

BabelNet is structured KB



(a) Excerpt of the WordNet graph centered on the synset play^1_n.



(b) Excerpt of the Wikipedia graph centered on the Wikipage Play (theatre).

- Both **WordNet** and **Wikipedia** can be taken as **knowledge graphs**.
- For WordNet, **nodes** are **synsets** and **edges** lexical and **semantic relations** between **synsets**.
- For Wikipedia, **nodes** are **Wikipages** and **edges** the **hyperlinks.**
- For **BabelNet**, these graphs are **combined**.

Parallel Corpora

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Idea: Parallel (sentence-aligned) data **implicitly** encodes multilingual knowledge

That is almost a personal record for me this autumn.

Das ist für mich fast ein persönlicher Rekord in diesem Herbst

è quasi il mio record personal dell' autunno

es la major marca que he alcanzado este otoño

Parallel Corpora

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Hypothesis: data contains all we need to learn relations between languages:

Peter has a black dog.
The dog's name is Bruno.
Peter walks him three times a day.
He also owns a black cat

Peter ana mbwa mweusi.
Jina la mbwa ni Bruno.
Peter anatembea naye mara tatu kwa siku.
Pia anamiliki paka mweusi.

Can you guess which words mean *dog* and *black* in Swahili?

Parallel Corpora - Examples

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Europarl corpus

- proceedings of the Europen Parliament from 1996, currently 21 languages, over 2 million sentences in the original 11 languages

Hansard French/English corpus:

- parallel texts in English and Canadian French, proceedings of the Canadian Parliament. 1970s to 80s.

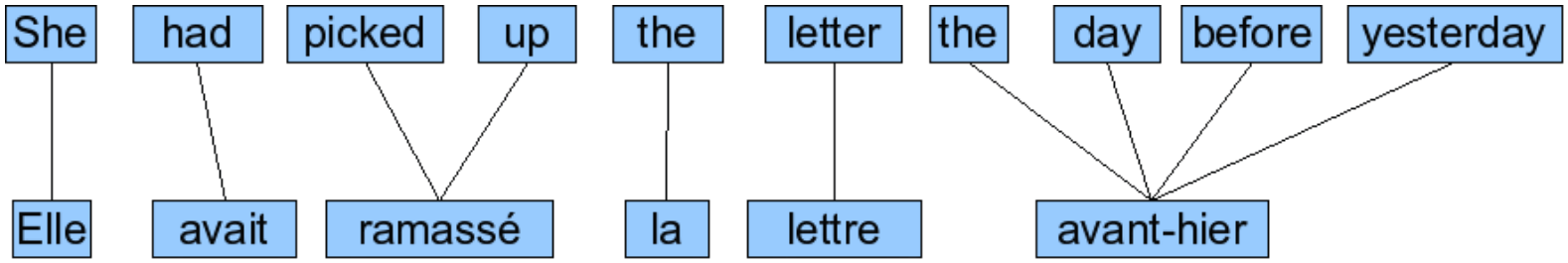
JRC-Acquis Multilingual Parallel Corpus

- EU legislation texts in 20 languages

How to use parallel data?

Statistical MT in a Nutshell:

1. Get a parallel corpus
2. (Align sentences)
3. Align words (based on probabilities, can be complex problem)
4. Merge words to phrases
5. Learn correspondences between phrases

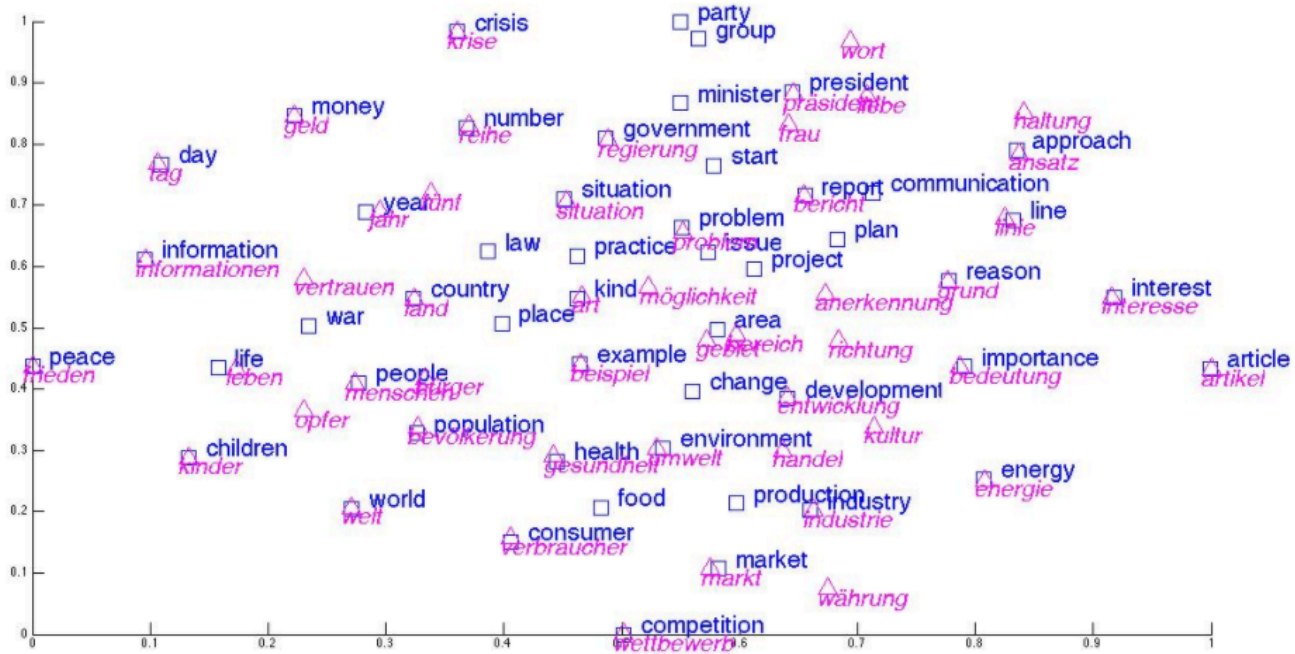


Connectionist Multilingual KBs

Multilingual Word Embeddings

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken



- **Goal:** represent multiple languages in the same embedding space
- Words with similar meaning should stay close to each other, `year` and `Jahr` have the same semantics and are similar to both `day` and `Tag`.
- Embeddings are learned from word co-occurrences.

Multilingual Word Embeddings

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- **What can we do with them?** – Basically, anything that we can do with monolingual embeddings but across languages:
  - Similarity between words
  - Aggregate them to compute similarity between sentences, documents…
  - Use them in applications that need relations between languages
    - Machine translation

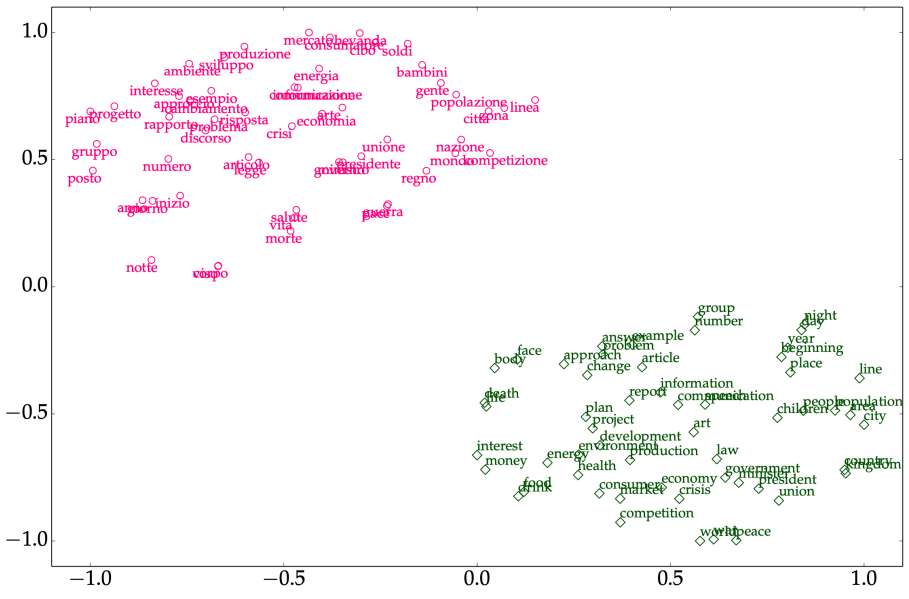
Multilingual Word Embeddings – How to get them

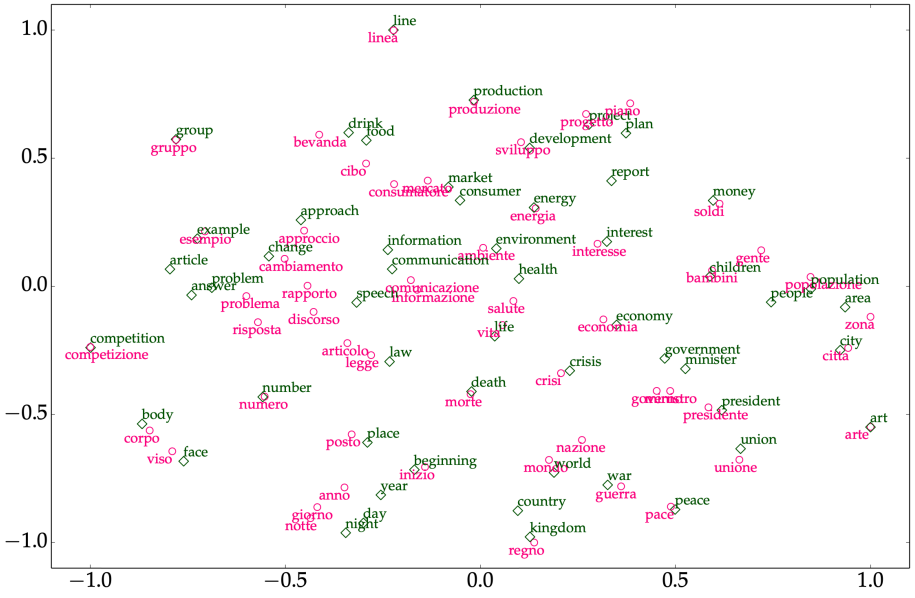
UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Monolingual mapping

- train one embedding space for each language,
- align the embedding spaces, by using lists of word pairs as anchors





Multilingual Word Embeddings – How to get them

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Learn from pseudo-crosslingual data

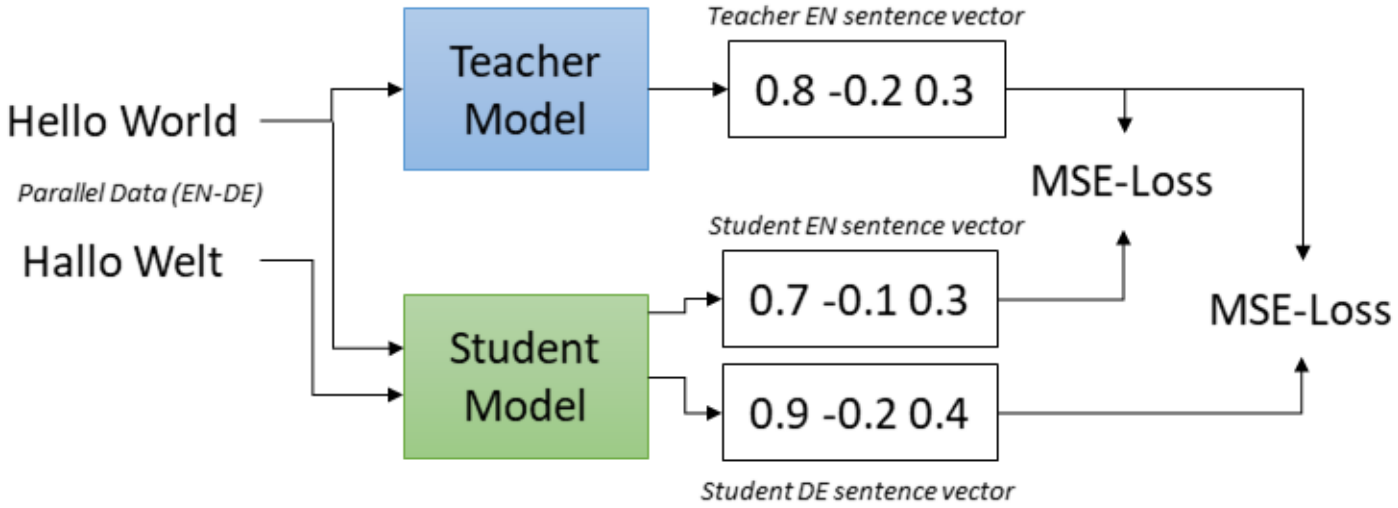
- Identify translation pairs such as `dog – Hund.`
- Replace 50% of all occurances of `dog` in the English corpus with `Hund` and the other way round for German.
- Concatenate the two corpora and learn embeddings.

Multilingual Word Embeddings – How to get them

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Use a corpus of sentence pairs in two languages.
- Integrate into the embedding training process the objective, that the two sentence must be similar.



XLM

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Use transformer architecture like BERT
- pre-trained for
  - Next word prediction objective —> on monolingual data
  - Masked language model —> on monolingual data
  - Translation language modeling —> bilingual data
    - an extension of MLM,
    - instead of considering monolingual text streams, concatenate parallel sentences
    - randomly mask words in both the source and target sentences.
    - To predict a word masked in an English sentence, the model can either attend to surrounding English words or to the German translation, encouraging the model to align the English and German representations.
    - In particular, the model can leverage the German context if the English one is not sufficient to infer the masked English words.
- Data: texts in 100 languages

Summary

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- **Multilingual KBs**
- **Examples of symbolic multilingual KB**
  - BabelNet
  - Parallel corpora
- **Examples of connectionist multilingual KBs**
  - Multilingual word embeddings

Reading Materials

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- **Mandatory**
  - Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217-250.
- **Optional**
  - Koehn, P. (2005, September). Europarl: A parallel corpus for statistical machine translation. In *MT summit* (Vol. 5, pp. 79-86).
  - Chen, P. J., Shen, J., Le, M., Chaudhary, V., El-Kishky, A., Wenzek, G., ... & Ranzato, M. A. (2019). Facebook AI's WAT19 Myanmar-English Translation Task Submission. *arXiv preprint arXiv:1910.06848*.
  - Schuster, S., Gupta, S., Shah, R., & Lewis, M. (2018). Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*.

Thank You