A Large-Scale Multilingual Disambiguation of Glosses
José Camacho Collados, Claudio Delli Bovi, Alessandro Raganato, and Roberto Navigli

lcl.uniroma1.it/disambiguated-glosses
definition

/nə'fɪ̱nən/  noun

noun: definition; plural noun: definitions

1. a statement of the exact meaning of a word, especially in a dictionary.
   "a dictionary definition of the verb"

   • an exact statement or description of the nature, scope, or meaning of something.
     "our definition of what constitutes poetry"
   synonyms: meaning, denotation, sense; More
   interpretation, explanation, elucidation, explication, description, clarification, exposition, expounding, illustration;
   deciphering, decoding;
   statement/outline of meaning
   "there is no agreed definition of 'intelligence'"

   • the action or process of defining something.
     "a question of definition"
Definitional Knowledge in NLP

- Word Sense Disambiguation
- Taxonomy/Ontology Learning
- Information Extraction
- Plagiarism Detection
- Question Answering
- ...

Navigli and Velardi, 2005
Lesk, 1986
Banerjee and Pedersen, 2002
Agirre and Soroa, 2009
Fernandez-Ordonez et al., 2012
Chen et al., 2014
Franco-Salvador et al., 2016
Camacho-Collados et al., 2015
Velardi et al., 2013
Flati et al., 2014
Espinosa-Anke et al., 2016
Richardson et al., 1998
Hill et al., 2015
Delli Bovi et al., 2015
Definitions and glosses are everywhere!

WordNet/Open Multilingual WordNet
150k definitions in 5 languages

Wiktionary
285k definitions in 1 language

Wikidata
8M definitions in 255 languages

OmegaWiki
118k definitions in 89 languages

Wikipedia
>30M definitions in 264 languages

A Large-Scale Multilingual Disambiguation of Glosses
José Camacho Collados, Claudio Delli Bovi, Alessandro Raganato, and Roberto Navigli
Disambiguating glosses on a large scale

Our goal:

✓ Construct a large-scale, multilingual repository of glosses and definitions with sense annotations
Disambiguating glosses on a large scale

Our goal:
✓ Construct a large-scale, multilingual repository of glosses and definitions with sense annotations

- The largest multilingual encyclopedic dictionary and semantic network
- Merger of 13 different knowledge resources
- >35M definitions in >250 languages!

babelnet.org
Disambiguating glosses on a large scale

Our goal:
✓ Construct a **large-scale, multilingual** repository of glosses and definitions with **sense annotations**

- The largest multilingual encyclopedic dictionary and semantic network
- Merger of 13 different knowledge resources
- **>35M definitions in >250 languages!**

How?

babelnet.org
Disambiguating glosses on a large scale

Problem:

- Disambiguating definitions is **hard**!
Problem:

- Disambiguating definitions is **hard**!

  *Interchanging the positions of the king and a rook.*
Disambiguating glosses on a large scale

Problem:

- Disambiguating definitions is hard!

*Interchanging the positions of the king and a rook.*

Definition of “castling” in chess (WordNet)

- Multilingual WSD/EL based on BabelNet (Moro et al., 2014)
Disambiguating glosses on a large scale

Our goal:
✓ Construct a large-scale, multilingual repository of glosses and definitions with sense annotations

Problem:
○ Disambiguating definitions is hard!
   ➞ Short and concise, not enough context
Disambiguating glosses on a large scale

Our goal:

✓ Construct a large-scale, multilingual repository of glosses and definitions with sense annotations

Problem:

○ Disambiguating definitions is hard!
  ➞ Short and concise, not enough context

Intuition:

○ Use various definitions of the same concept or entity at the same time and in multiple languages
Step 1: Context-rich Disambiguation
Step 1: Context-rich Disambiguation

- **Multilingual preprocessing pipeline:**
  - **Tokenization:** from the Polyglot project (165 languages)
  - **Part-of-speech tagging:** Stanford parser trained on Universal Dependencies (30 languages)
Step 1: Context-rich Disambiguation

○ **Multilingual preprocessing pipeline:**
  - **Tokenization:** from the Polyglot project (165 languages)
  - **Part-of-speech tagging:** Stanford parser trained on Universal Dependencies (30 languages)

○ **Context enrichment:**
  - Given a *definiendum*, collect all its definitions in every available language and resource and bring them together into a *single, heterogeneous multilingual text!*
Step 1: Context-rich Disambiguation

- **Multilingual preprocessing pipeline:**
  - **Tokenization:** from the Polyglot project (165 languages)
  - **Part-of-speech tagging:** Stanford parser trained on Universal Dependencies (30 languages)

- **Context enrichment:**
  - Given a *definiendum*, collect all its definitions in every available language and resource and bring them together into a *single, heterogeneous multilingual text!*
Step 1: Context-rich Disambiguation

- Babelfy (Moro et al., 2014):
  - Unified graph-based approach to multilingual **Word Sense Disambiguation** and **Entity Linking**
  - Designed to handle multilingual text ("language-agnostic" setting)
Step 1: Context-rich Disambiguation

Our running example: castling

Interchanging the positions of the **king** and a **rook**.

**Castling** is a move in the game of **chess** involving a player’s **king** and either of the player’s original **rooks**.

A move in which the **king** moves two **squares** towards a **rook**, and the **rook** moves to the other side of the **king**.
Step 1: Context-rich Disambiguation

Our running example: castling

*Interchanging the positions of the *king* and a *rook*. 

*Castling* is a move in the game of *chess* involving a player's *king* and either of the player's original *rooks*. 

A move in which the *king* moves two *squares* towards a *rook*, and the *rook* moves to the other side of the *king*.
Step 1: Context-rich Disambiguation

Our running example: castling

Interchanging the positions of the king and a rook.

Castling is a move in the game of chess involving a player’s king and either of the player’s rooks.

Manœuvre du jeu d’échecs

Manöver im Schach

Manøver i sjakk

Manovra nel gioco degli scacchi

El enroque es un movimiento especial en el juego de ajedrez que involucra al rey y a una de las torres del jugador.

Rošáda je zvláštní tah v šachu, při kterém táhne zároveň král a věž.

Rokade er et spesialtrekk i sjakk.

Rokade er et spesialtrekk i sjakk.

Rokade er et spesialtrekk i sjakk.

A Large-Scale Multilingual Disambiguation of Glosses
José Camacho Collados, Claudio Delli Bovi, Alessandro Raganato, and Roberto Navigli
Step 2: Disambiguation Refinement

Our running example: castling

Interchanging the positions of the **king** and a **rook**.

**Castling** is a move in the game of **chess** involving a player’s **king** and either of the player's original **rooks**.

A move in which the **king** moves two **squares** towards a **rook**, and the **rook** moves to the other side of the **king**.

**El enroque** es un movimiento especial en el juego de **ajedrez** que involucra al **rey** y a una de las **torres** del jugador.

**Rokade** er et spesialtrekk i **sjakk**.

**Rošáda** je zvláštní tah v šachu, při kterém táhne zároveň král a věž.

**Spielzug im Schach**, bei dem **König** und Turm einer Farbe bewegt werden.
Step 2: Disambiguation Refinement

**SEMANTIC SIMILARITY**

**NASARI_embed**: Latent semantic representations of BabelNet synsets and Wikipedia pages as 300-dimensional vectors.

- **Goal**: Re-disambiguate low confidence annotations from the first step.

- **How**: We obtain the centroid NASARI vector of high-confidence annotations and compute cosine similarity with all the candidate synsets NASARI vectors.

(Camacho-Collados, Pilehvar and Navigli, ACL 2015)
Step 2: Disambiguation Refinement

Our running example: castling

*Interchanging the positions of the king and a rook.*

Castling is a move in the game of chess involving a player's king and either of the player's rooks.

Manoeuvre du jeu d'échecs

Manöver im Schach

Rošáda je zvláštní tah v šachu, při kterém táhne zároveň král a věž.

El enroque es un movimiento especial en el juego de ajedrez que involucra al rey y a una de las torres del jugador.

Rokade er et spesialtrekk i sjakk.

Rokade er et spesialtrekk i sjakk.

To rokke είναι μια ειδική κίνηση στο σκάκι που συμμετέχουν ο βασιλιάς και ένας από τους δύο πύργους.
Step 2: Disambiguation Refinement

**Our running example: castling**

*Interchanging the positions of the king and a rook.*

**Castling** is a move in the game of chess involving a player’s king and either of the player’s rooks.

**Our running example: castling**


In English, castling is a move in the game of chess in which the king moves two squares towards a rook, and the rook moves to the other side of the king.
Step 2: Disambiguation Refinement

Our running example: castling

Interchanging the positions of the **king** and a **rook**.
Step 2: Disambiguation Refinement

Our running example: castling

Interchanging the positions of the king and a rook.

Castling is a move in the game of chess involving a player's king and either of the player's original rooks.

El enroque es un movimiento especial en el juego de ajedrez que involucra al rey y a una de las torres del jugador.

Rošáda je zvláštní tah v šachu, při kterém táhne zároveň král a věž.

Rokade er et spesialtrekk i sjakk.

To roké éíni mia eidikí kívnos sto skáki pou summetéchoun o basiliáas kai énas apó tous duo purgous.
Step 2: Disambiguation Refinement

Our running example: castling

Interchanging the positions of the king and a rook.

Castling is a move in the game of chess involving a player's king and either of the player's original rooks.

A move in which the king moves two squares towards a rook, and the rook moves to the other side of the king.
Evaluation

- **Extrinsic Evaluation:**
  - Open Information Extraction (DefIE)
  - Sense Clustering (NASARI)

- **Manual Intrinsic Evaluation:**
  - 3 languages (EN, IT, ES)
  - Sample of 100 definitions each
Extrinsic Evaluation I: Open Information Extraction

Large-Scale Information Extraction from Textual Definitions through Deep Syntactic and Semantic Analysis (Delli Bovi et al., TACL 2015)

DefIE

- DefIE uses disambiguated definitions as input. We simply plugged-in our disambiguated definitions as input and leave its whole pipeline unchanged.

- This leaves to improvements according to both manual and automatic evaluation

lcl.uniroma1.it/defie/
Extrinsic Evaluation I: Open Information Extraction

Evaluation on a sample of 150 definitions

NUMBER OF EXTRACTIONS

PRECISION

A Large-Scale Multilingual Disambiguation of Glosses
José Camacho Collados, Claudio Delli Bovi, Alessandro Raganato, and Roberto Navigli
Extrinsic Evaluation II: Construction of NASARI+

NASARI semantic representation construction pipeline (ACL 2015)
Extrinsic Evaluation II: Construction of NASARI+

We simply enrich BabelNet taxonomy with the high-precision disambiguated glosses. The whole pipeline remains unchanged.
Extrinsic Evaluation II: Wikipedia Sense Clustering

ACCURACY

A Large-Scale Multilingual Disambiguation of Glosses
José Camacho Collados, Claudio Delli Bovi, Alessandro Raganato, and Roberto Navigli
Intrinsic Evaluation

Manual evaluation on a sample of 300 definitions

PRECISION OF THE THREE DIFFERENT DISAMBIGUATION STRATEGIES
Intrinsic Evaluation

Manual evaluation on a sample of 300 definitions

Coverage of the high-precision version of the corpus: ~65% for all PoS and ~75% for nouns

Precision of the three different disambiguation strategies

A Large-Scale Multilingual Disambiguation of Glosses
José Camacho Collados, Claudio Delli Bovi, Alessandro Raganato, and Roberto Navigli
Overview of the release

- Two different versions of the corpus:
  - Complete version before refinement (Step 1)
  - High-Precision version after refinement (Step 2)

- Formatted in an easy-to-process XML, divided by language and resource
Overview of the release

- Two different versions of the corpus:
  - Complete version before refinement (Step 1)
Overview of the release

- Two different versions of the corpus:
  - Complete version before refinement (Step 1)

```xml
<definition resource="WIKI" id="Palaeochiropteryx">
   <text>Palaeochiropteryx is an extinct genus of bat from the Middle Eocene</text>
   <annotations>
      <annotation source="BABELFY" anchor="Palaeochiropteryx" bfScore="0.9330" coherence="1.0000" />
      <annotation source="MCS" anchor="extinct" bfScore="--" coherence="--" />
      <annotation source="BABELFY" anchor="genus" bfScore="0.9231" coherence="1.0000" />
      <annotation source="BABELFY" anchor="bat" bfScore="0.9231" coherence="1.0000" />
      <annotation source="MCS" anchor="Middle" bfScore="--" coherence="--" />
      <annotation source="BABELFY" anchor="Middle Eocene" bfScore="1.0000" coherence="1.0000" />
      <annotation source="BABELFY" anchor="Eocene" bfScore="1.0000" coherence="1.0000" />
      <annotation source="BABELFY" anchor="Europe" bfScore="0.8586" coherence="1.0000" />
   </annotations>
</definition>
```
Overview of the release

- Two different versions of the corpus:
  - Complete version before refinement (Step 1)
Overview of the release

- Two different versions of the corpus:
  - Complete version before refinement (Step 1)
  - High-Precision version after refinement (Step 2)
Statistics - #Sense annotations

Before refinement (Step 1)

All languages

- Babelfy: 69.8%
- MCS: 30.2%

249,544,708 annotations

After refinement (Step 2)

All languages

- Babelfy: 88.7%
- NASARI: 11.3%

163,029,131 annotations
Statistics - #Sense annotations per language

- English: ~58.8M (~37.9M before refinement, ~14.1M after refinement)
- Spanish: ~14.1M (~8.3M before refinement, ~5.2M after refinement)
- French: ~18.2M (~10.6M before refinement, ~8.4M after refinement)
- Italian: ~13.4M (~8.4M before refinement, ~5.0M after refinement)
- Persian: ~18.2M (~13.4M before refinement, ~4.8M after refinement)
Conclusion

A large-scale multilingual corpus of disambiguated glosses:

- **250 milion sense-annotations** for both concepts and named entities
- In total, over **35 million definitions** have been disambiguated
- **256 languages**
- Both versions of the corpus freely available online

PLAY WITH ME!

http://lcl.uniroma1.it/disambiguated-glosses/
Thank you!

http://lcl.uniroma1.it/disambiguated-glosses/