A Large-Scale Multilingual Disambiguation of Glosses

José Camacho Collados, Claudio Delli Bovi,
Alessandro Raganato, and Roberto Navigli

Dipartimento
di Informatica

Sapienza
Università di Roma

lcl.uniroma1.it/disambiguated-glosses

definition

/dɛfɪˈnɪʃ(ə)n/

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

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**

BabelNet

babelnet.org

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

Disambiguating glosses on a large scale

Our goal:

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

babelnet.org

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

How?

Disambiguating glosses on a large scale

Problem:

- Disambiguating definitions is **hard**!

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)

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)

Babelfy

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)

Today at **LREC**, **Session O19**!

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)

Babelfy

- **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)

Babelfy

babelfy.org

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.***

*Manœuvre du jeu d'**échecs***

***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*

*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**.*

***Rok** İngilizce'de kaleye **rook** denmektedir.*

Το ροκέ είναι μια ειδική **κίνηση** στο **σκάκι** που συμμετέχουν ο βασιλιάς και ένας από τους δυο **πύργους**.

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**.

which the **king** moves two ards a **rook**, and the **rook** other side of the **king.**

Man… d'**échec**…

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

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**.

Rok İngilizce'de kaleye **rook** denmektedir.

Το ροκέ είναι μια ειδική **κίνηση** στο **σκάκι** που συμμετέχουν ο βασιλιάς και ένας από τους δυο **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 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**.*

*Manœuvre du jeu d'**échecs***

***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*

*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

***Rok** İngilizce'de kaleye **rook** denmektedir.*

Το ροκέ είναι μια ειδική **κίνηση** στο **σκάκι** που συμμετέχουν ο βασιλιάς και ένας από τους δυο **πύργους**.

Step 2: Disambiguation Refinement

SEMANTIC SIMILARITY

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

NASARI lcl.uniroma1.it/nasari/

(Camacho-Collados, Pilehvar and Navigli, ACL 2015)

- **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.

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**.*

*which the **king** moves two ards a **rook**, and the **rook** other side of the **king**.*

*Man d'**échec***

***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*

*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**.*

__Rok__ İngilizce'de kaleye __rook__ denmektedir.

Το ροκέ είναι μια ειδική **κίνηση** στο **σκάκι** που συμμετέχουν ο βασιλιάς και ένας από τους δυο **πύργους**.

Step 2: Disambiguation Refinement

Our running example: castling

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

V chess

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

V rook-chess

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

V chess

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

V rook-chess

Spielzug im ***Schach****, bei dem* ***König*** *und* ***Turm*** *einer Farbe bewegt werden*

V rook-chess

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

V chess

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

V chess

Το ροκέ είναι μια ειδική **κίνηση** στο **σκάκι** που συμμετέχουν ο βασιλιάς και ένας από τους δυο **πύργους**.

V chess

Rok *İngilizce'de kaleye* ***rook*** *denmektedir.*

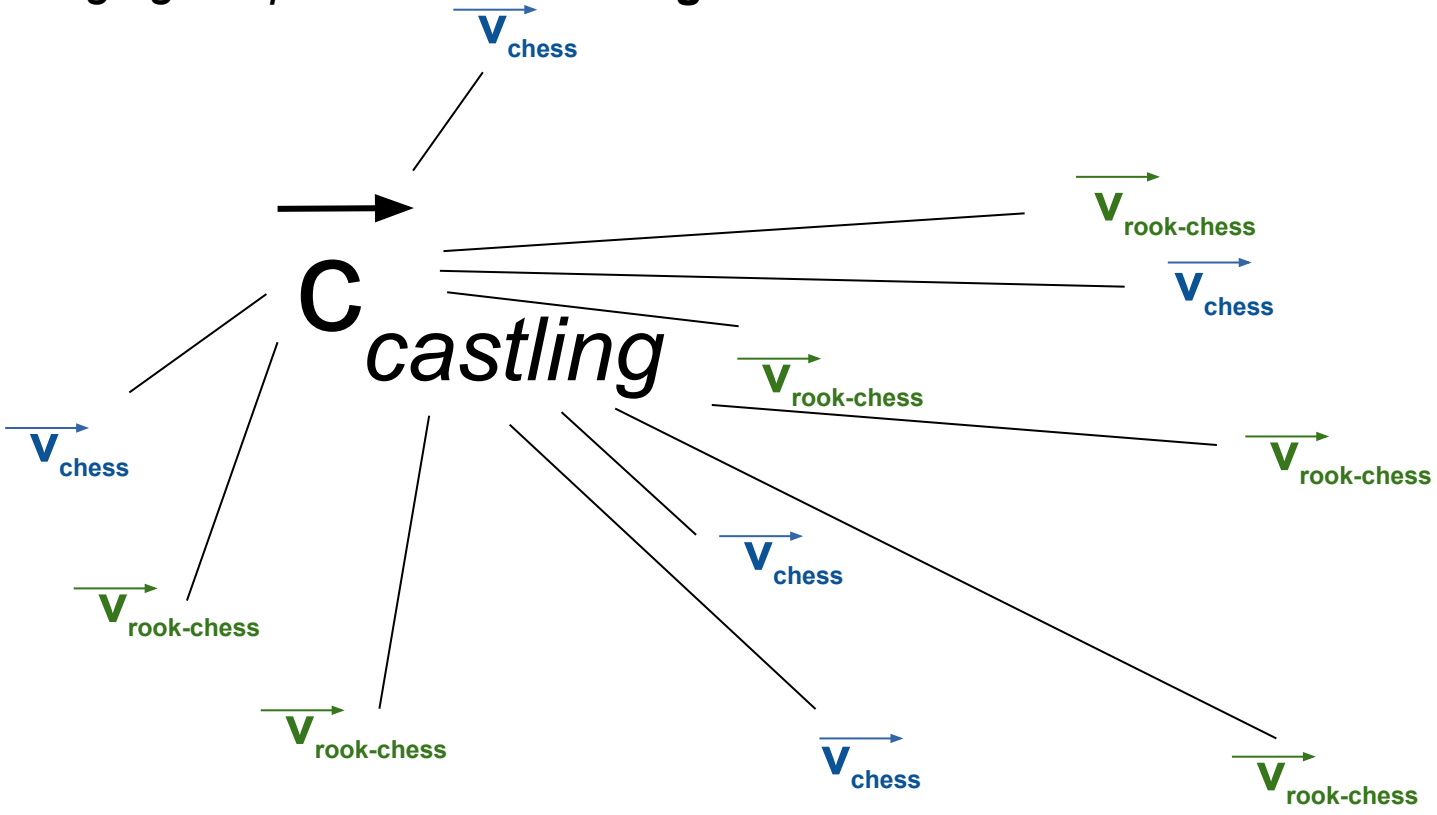
V rook-chess

V rook-chess

Step 2: Disambiguation Refinement

Our running example: castling

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



Step 2: Disambiguatio

Our running example: castlin

Interchanging the positions

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

ch the ***king*** *moves two* *rds a* ***rook****, and the* ***rook*** *moves to the other side of the* ***king****.*

*Manœuvre du jeu d'***échecs**

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*

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***

Rok *İngilizce'de kaleye* ***rook*** *denmektedir.*

Το ροκέ είναι μια ειδική ***κίνηση*** στο **σκάκι** που συμμετέχουν ο βασιλιάς και ένας από τους δυο **πύργους**.

Step 2: Disambiguatio

Our running example: castlin

Interchanging the positions

Castling is a move in the game of *ch the king moves two involving a player's king and either* *rds a rook, and the rook player's original rooks.* *ther side of the king.*

θ<0.5

Manœuvre du jeu d'échecs

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

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

Rokade er et spesialtrekk i sjakk

Rok İngilizce'de kale *ro* *denmektedir.*

θ=0.86

Το ροκέ είναι μια ειδική κίνηση στο **σκάκι** που συμμετέχουν ο βασιλιάς και ένας από τους δυο **πύργους**.

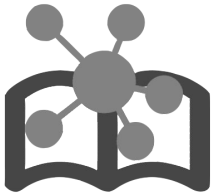
Evaluation

- **Extrinsic Evaluation:**
  - Open Information Extraction (**DeflE**)
  - 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 

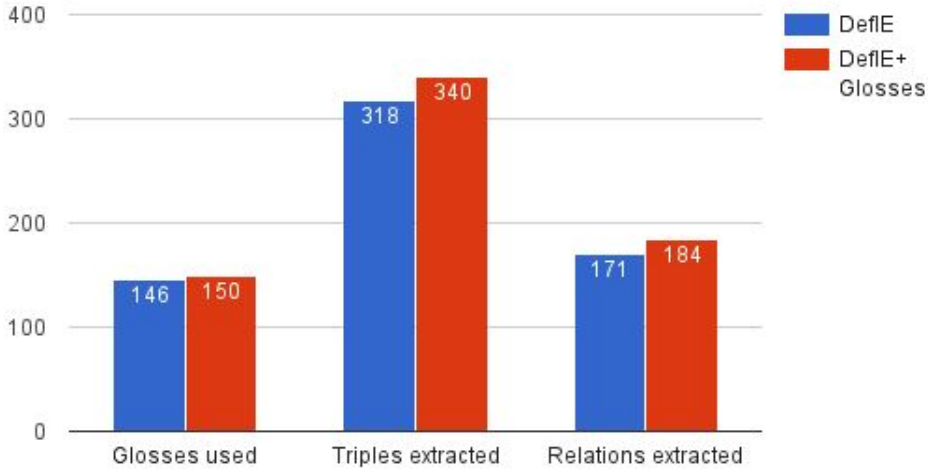
lcl.uniroma1.it/defie/

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

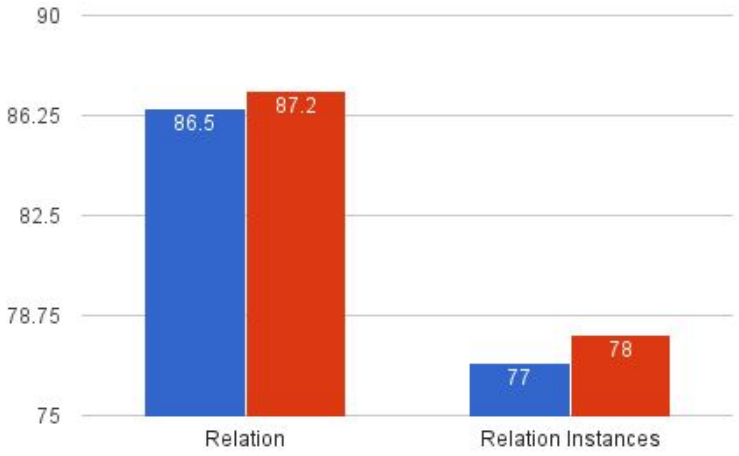
- This leaves to improvements according to both manual and automatic evaluation

Extrinsic Evaluation I: Open Information Extraction

Evaluation on a sample of 150 definitions



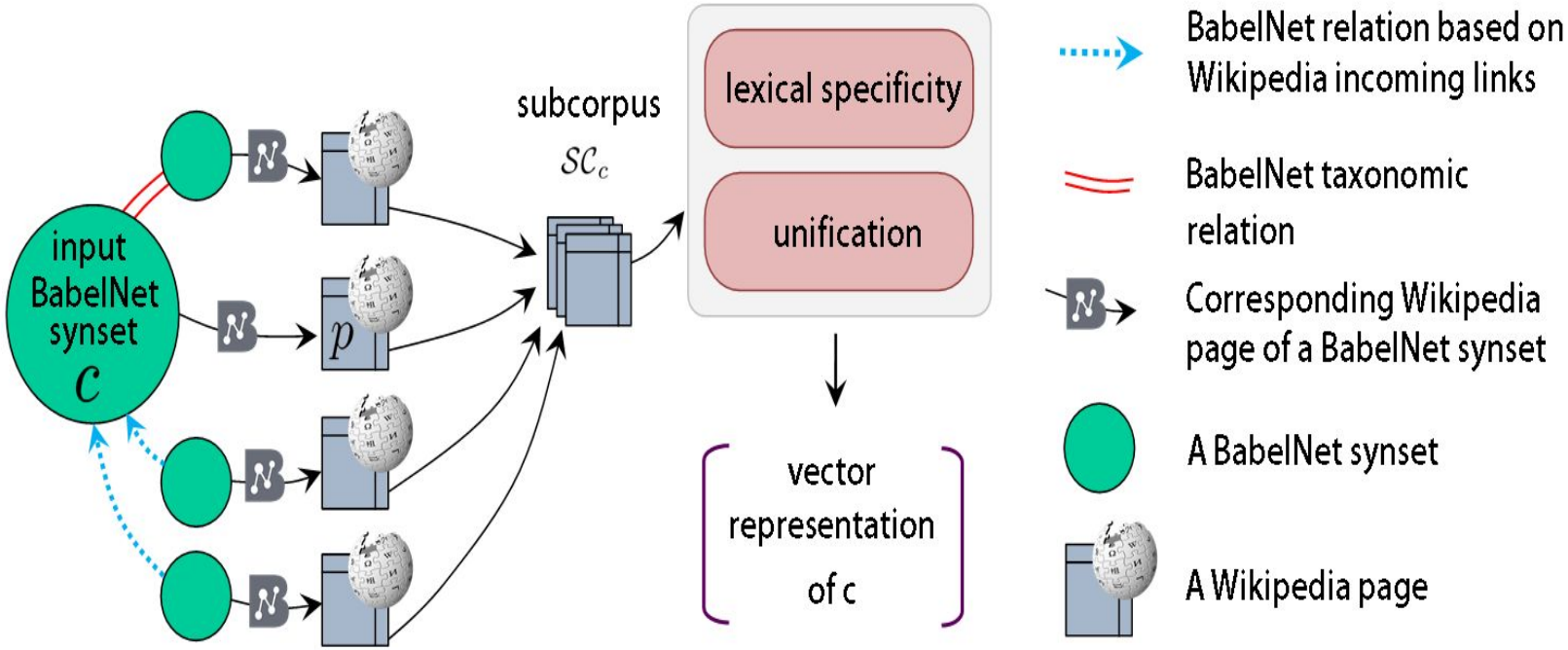
NUMBER OF EXTRACTIONS



PRECISION

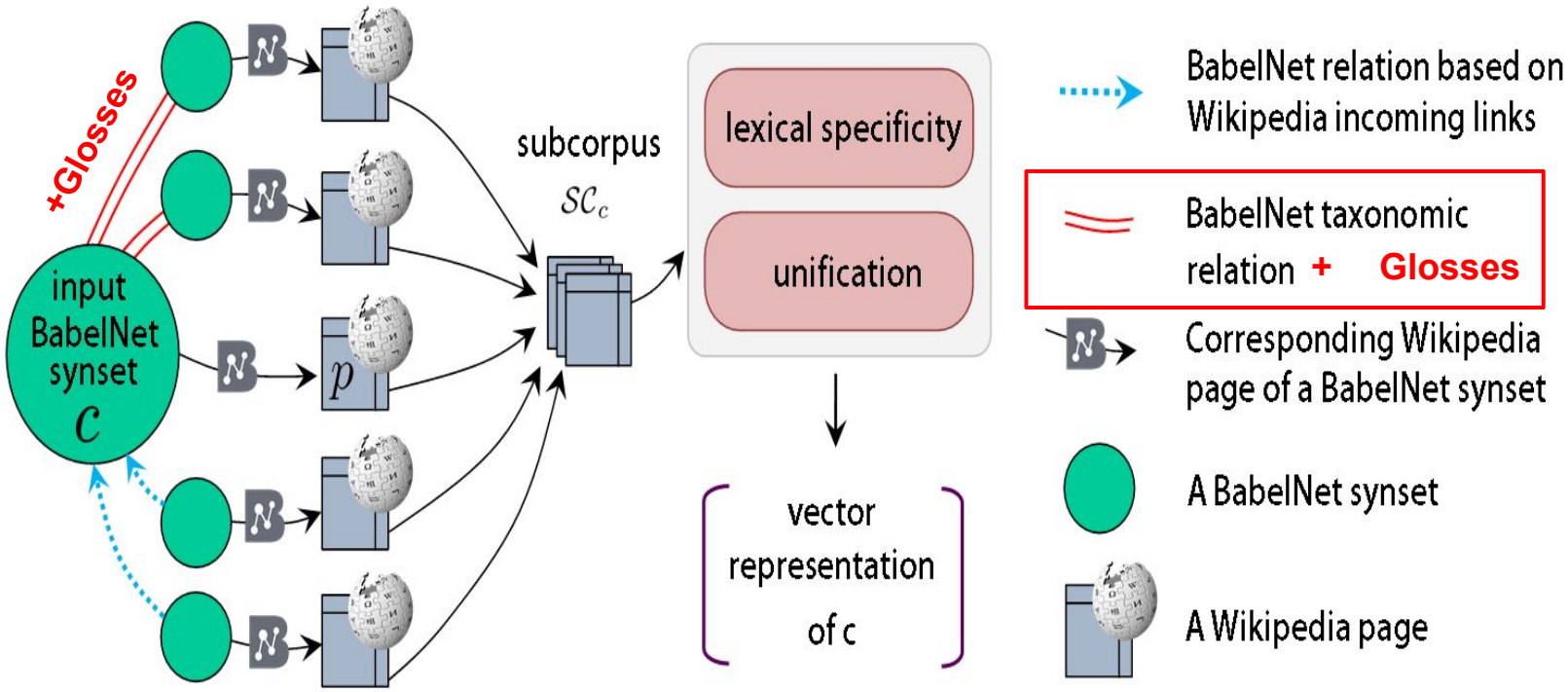
Extrinsic Evaluation II: Construction of NASARI+

NASARI semantic representation construction pipeline (ACL 2015)

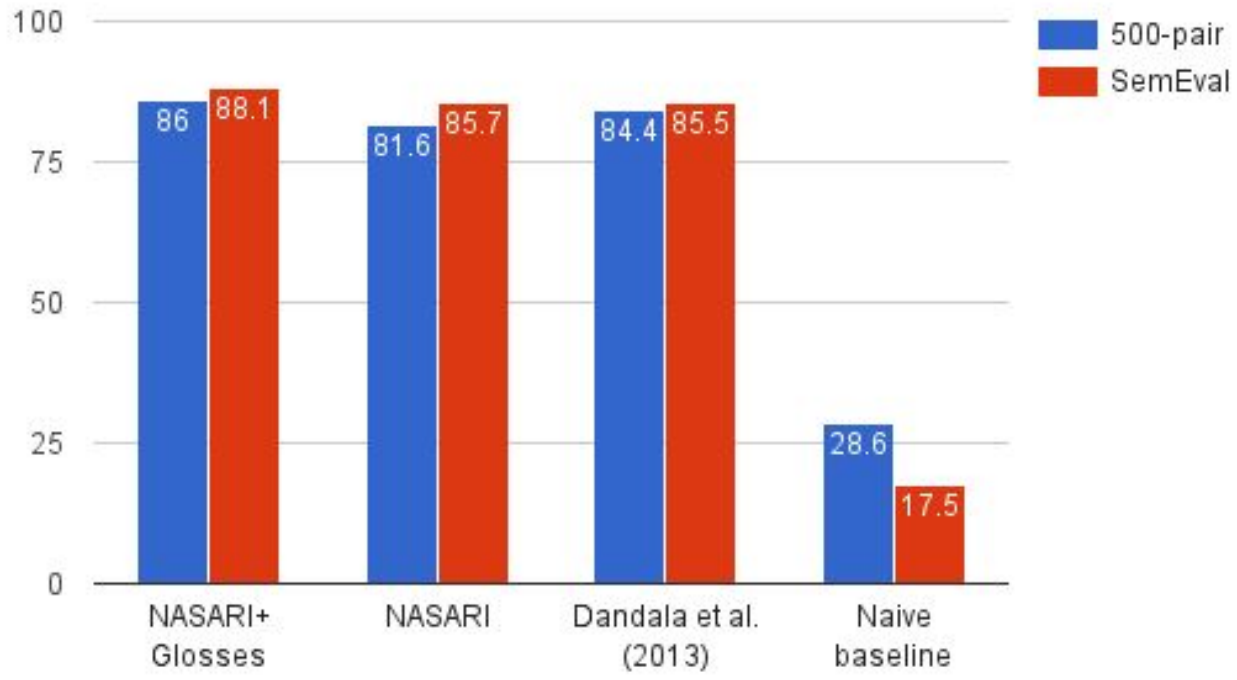


Extrinsic Evaluation II: Construction of NASARI+

We simply enrich BabelNet taxonomy with the high-precision disambiguated glosses. The whole pipeline remains unchanged.



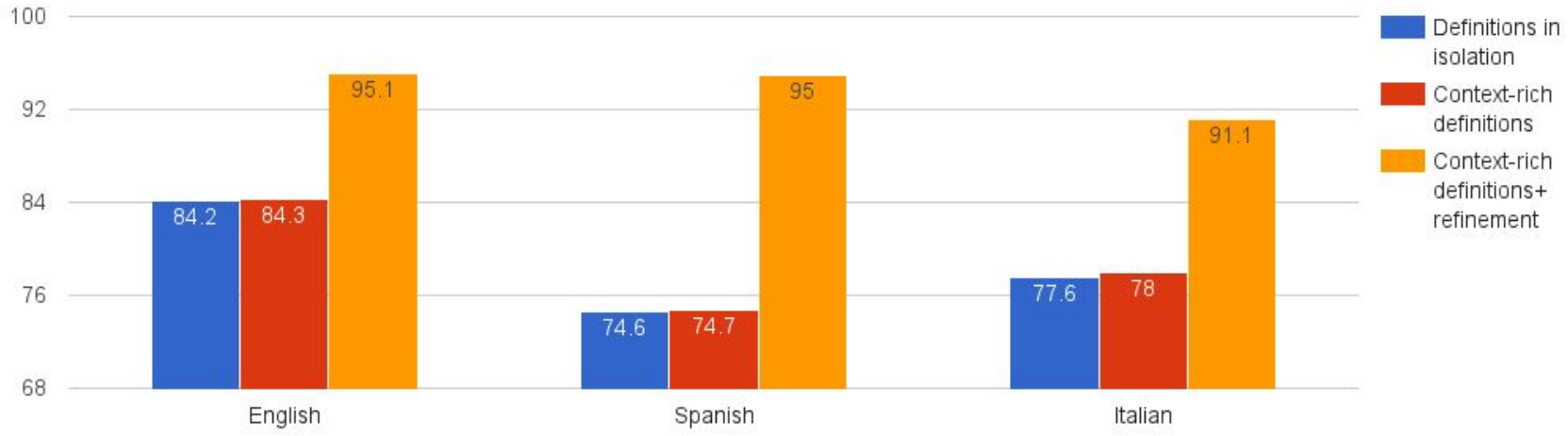
Extrinsic Evaluation II: Wikipedia Sense Clustering



ACCURACY

Intrinsic Evaluation

Manual evaluation on a sample of 300 definitions

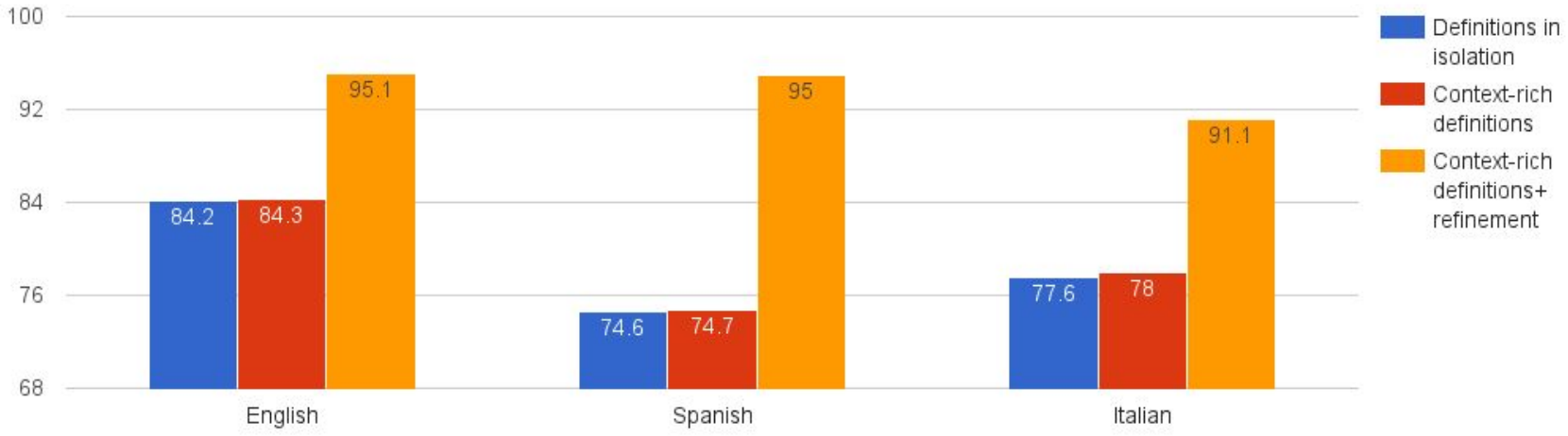


PRECISION OF THE THREE DIFFERENT DISAMBIGUATION STRATEGIES

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

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)

```
<definition resource="WIKI" id="Palaeochiropteryx">
        <text>Palaeochiropteryx is an extinct genus of bat from the Middle Eocene of Europe.</text>
        <annotations>
                <annotation source="BABELFY" anchor="Palaeochiropteryx" bfScore="1.0000" coherenceScore="0.4722">bn:03408711n</annotation>
                <annotation source="MCS" anchor="extinct" bfScore="--" coherenceScore="--">bn:00102630a</annotation>
                <annotation source="BABELFY" anchor="genus" bfScore="0.9330" coherenceScore="0.6944">bn:00037780n</annotation>
                <annotation source="BABELFY" anchor="bat" bfScore="0.9231" coherenceScore="0.6667">bn:00008977n</annotation>
                <annotation source="MCS" anchor="Middle" bfScore="--" coherenceScore="--">bn:00017120n</annotation>
                <annotation source="BABELFY" anchor="Middle Eocene" bfScore="1.0000" coherenceScore="0.6111">bn:00031088n</annotation>
                <annotation source="BABELFY" anchor="Eocene" bfScore="1.0000" coherenceScore="0.6111">bn:00031088n</annotation>
                <annotation source="BABELFY" anchor="Europe" bfScore="0.8586" coherenceScore="0.8333">bn:00031896n</annotation>
        </annotations>
</definition>
```

Overview of the release

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

```
<definition resource="WIKI" id="Palaeochiropteryx">
        <text>Palaeochiropteryx is an extinct genus of bat from the Middle Eoce
        <annotations>
                <annotation source="BABELFY" anchor="Palaeochiropteryx" bfScore
                <annotation source="MCS" anchor="extinct" bfScore="--" coherenc
                <annotation source="BABELFY" anchor="genus" bfScore="0.9330" co
                <annotation source="BABELFY" anchor="bat" bfScore="0.9231" cohe
                <annotation source="MCS" anchor="Middle" bfScore="--" coherence
                <annotation source="BABELFY" anchor="Middle Eocene" bfScore="1.
                <annotation source="BABELFY" anchor="Eocene" bfScore="1.0000" c
                <annotation source="BABELFY" anchor="Europe" bfScore="0.8586" c
        </annotations>
</definition>
```

Overview of the release

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

```
>
s of bat from the Middle Eocene of Europe.</text>

r="Palaeochiropteryx" bfScore="1.0000" coherenceScore="0.4722">bn:03408711n</annotation>
xtinct" bfScore="--" coherenceScore="--">bn:00102630a</annotation>
r="genus" bfScore="0.9330" coherenceScore="0.6944">bn:00037780n</annotation>
r="bat" bfScore="0.9231" coherenceScore="0.6667">bn:00008977n</annotation>
iddle" bfScore="--" coherenceScore="--">bn:00017120n</annotation>
r="Middle Eocene" bfScore="1.0000" coherenceScore="0.6111">bn:00031088n</annotation>
r="Eocene" bfScore="1.0000" coherenceScore="0.6111">bn:00031088n</annotation>
r="Europe" bfScore="0.8586" coherenceScore="0.8333">bn:00031896n</annotation>
```

Overview of the release

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

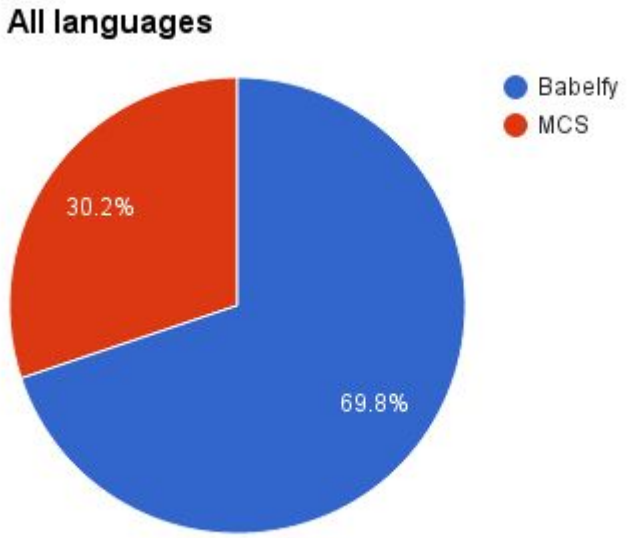
```
<definition resource="WIKI" id="Palaeochiropteryx">
        <text>Palaeochiropteryx is an extinct genus of bat from the Middle Eocene of Europe.</text>
        <annotations>
                <annotation source="BABELFY" anchor="Palaeochiropteryx" bfScore="1.0000" coherenceScore="0.4722">bn:03408711n</annotation>
                <annotation source="MCS" anchor="extinct" bfScore="--" coherenceScore="--">bn:00102630a</annotation>
                <annotation source="BABELFY" anchor="genus" bfScore="0.9330" coherenceScore="0.6944">bn:00037780n</annotation>
                <annotation source="BABELFY" anchor="bat" bfScore="0.9231" coherenceScore="0.6667">bn:00008977n</annotation>
                <annotation source="MCS" anchor="Middle" bfScore="--" coherenceScore="--">bn:00017120n</annotation>
                <annotation source="BABELFY" anchor="Middle Eocene" bfScore="1.0000" coherenceScore="0.6111">bn:00031088n</annotation>
                <annotation source="BABELFY" anchor="Eocene" bfScore="1.0000" coherenceScore="0.6111">bn:00031088n</annotation>
                <annotation source="BABELFY" anchor="Europe" bfScore="0.8586" coherenceScore="0.8333">bn:00031896n</annotation>
        </annotations>
</definition>
```

  - High-Precision version after refinement (Step 2)

```
<definition resource="WN" id="11132462n">
        <text>16th President of the United States; saved the Union during the American Civil War and emancipated the slaves; was assassinated by Booth (1809-1865)</text>
        <annotations>
                <annotation source="NASARI" anchor="President" bfScore="--" coherenceScore="--" nasariScore="0.7639">bn:00018323n</annotation>
                <annotation source="BABELFY" anchor="President of the United States" bfScore="0.7852" coherenceScore="0.6272" nasariScore="0.7639">bn:00018323n</annotation>
                <annotation source="BABELFY" anchor="United States" bfScore="0.7459" coherenceScore="0.9084" nasariScore="0.7263">bn:00003341n</annotation>
                <annotation source="BABELFY" anchor="States" bfScore="0.9889" coherenceScore="0.9084" nasariScore="0.7263">bn:00003341n</annotation>
                <annotation source="BABELFY" anchor="16th President of the United States" bfScore="1.0000" coherenceScore="0.6572" nasariScore="0.8042">bn:00000388n</annotation>
        </annotations>
</definition>
```

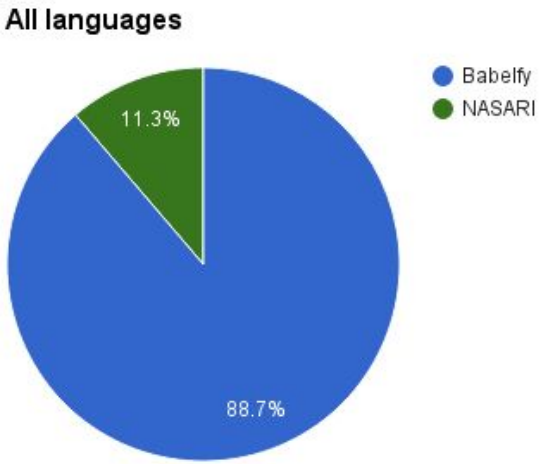
Statistics - #Sense annotations

Before refinement (Step 1)



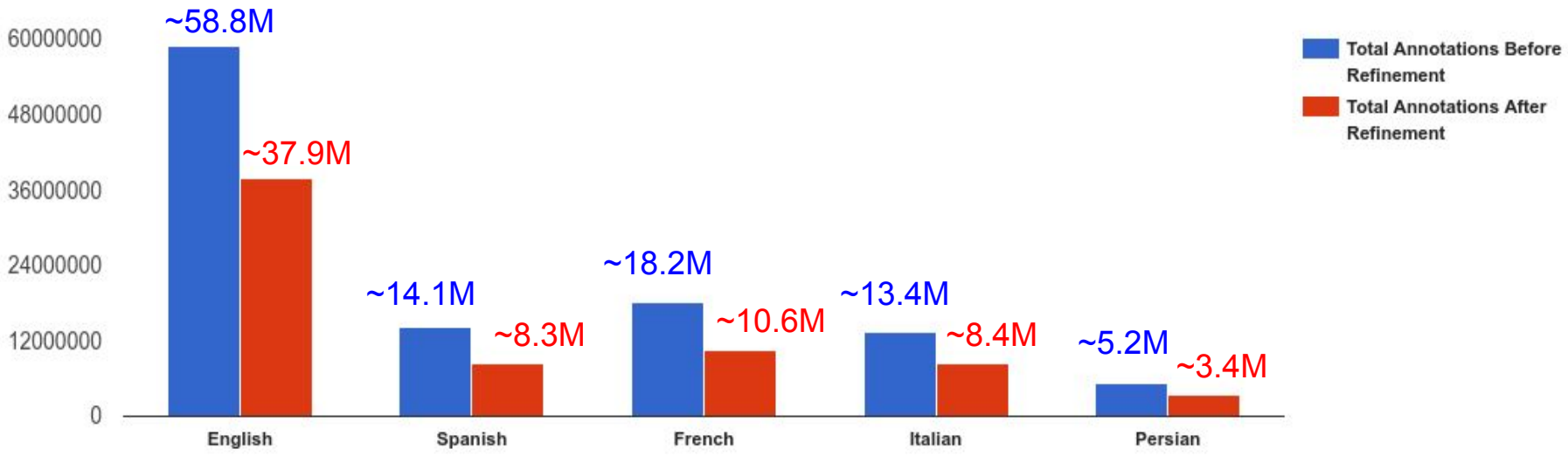
249,544,708 annotations

After refinement (Step 2)



163,029,131 annotations

Statistics - #Sense annotations per language



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/