Named Entity Recognition and Linking

Artūrs Znotiņš, arturs.znotins@gmail.com

Supervisor: Guntis Bārzdiņš, Dr.sc.comp.

The Problem

- Large amounts of unstructured data – natural language (WWW, books, newspapers, radio)
- A lot of ambiguity - context is very important
- Humans are good at semantic disambiguation:
  - **What entities does text refer to?**
  - What facts does text describe?
  - What is the meaning of specific word?
- How to do this automatically?

Motivation

- Utilize existing unstructured natural language resources

  Hidden knowledge

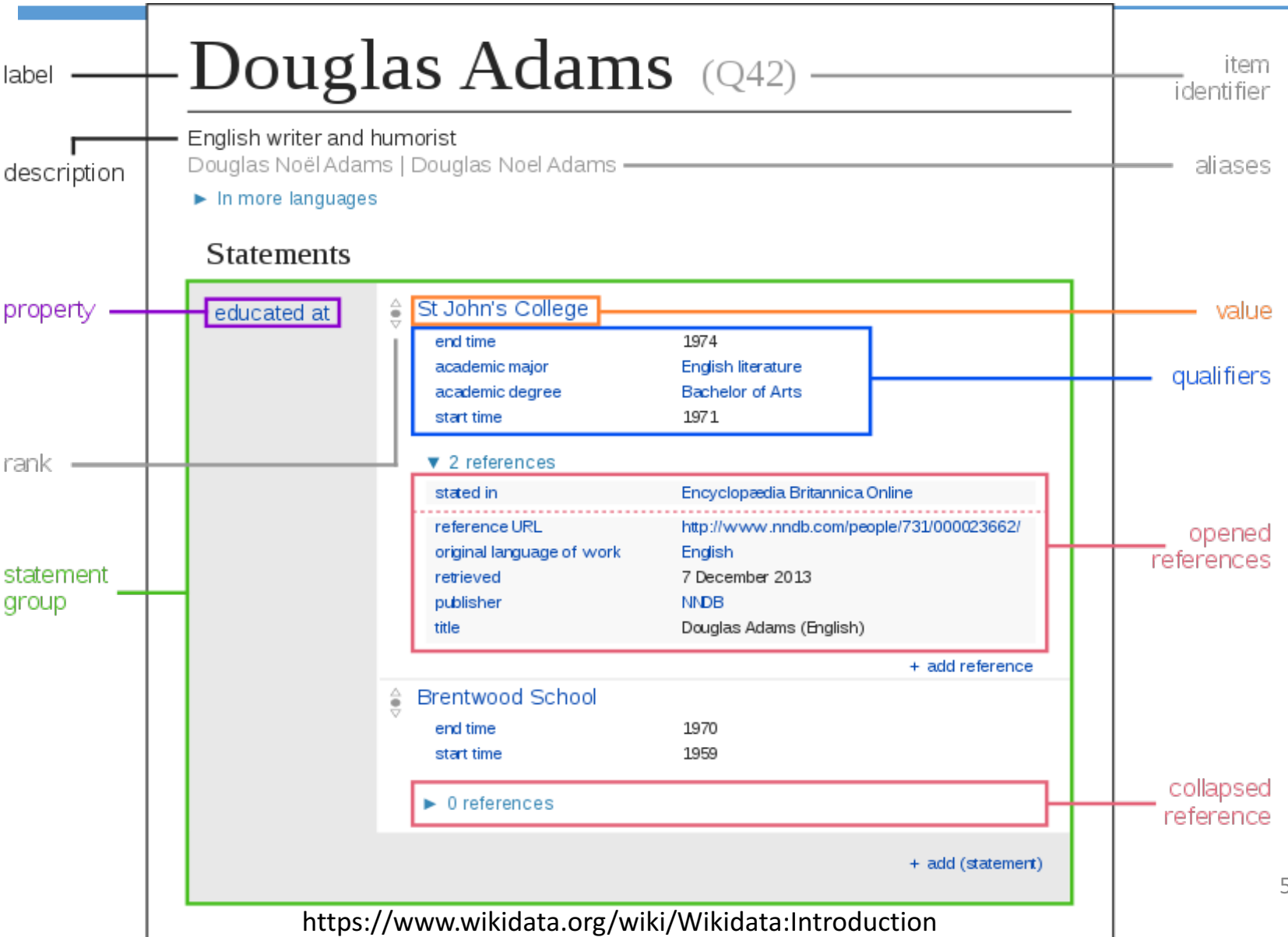
- Use cases:
  - Information extraction (IE)
  - Text clustering
  - Media monitoring
  - Dialog systems
  - Question answering
  - Machine translation

Cikos atiet vilciens uz Daugavpili?

Definitions

- **Entity**: something that exists by itself
- **Named entity (NE):** entity with specific name
- **Mention:** phrase that refers to entity
- **Knowledge base (KB)**: organized repository of knowledge consisting of concepts, properties, links
- **(Named) Entity Linking (NEL)**: linking mentions of entities within a text to KB entities
- **Word Sense Disambiguation (WSD)**: assigning meanings to word occurrences within text

Knowledge Base: Example



https://www.wikidata.org/wiki/Wikidata:Introduction

Knowledge Base: Example

Andris Bērziņš (Q3744607): Latvian politician
15 KB (62 words) - 10:55, 19 September 2017

Andris Bērziņš (Q57506): eighth president of Latvia
39 KB (348 words) - 10:47, 29 October 2017

Andris Bērziņš (Q380986): Prime Minister of Latvia
23 KB (204 words) - 15:40, 25 November 2017

Andris Bērziņš (Q238915): Wikimedia disambiguation page
9 KB (277 words) - 21:59, 11 November 2017

Andris Bērziņš (Q10863065): Latvian actor
11 KB (76 words) - 10:34, 19 September 2017

https://www.wikidata.org/w/index.php?search=&search=andris+berzins

Named Entity Linking: Example

*"Arī **otra figūra** Daimler lietā ir **Bojāra** ārštata **padomnieks**, un sens **eksmēra draugs** no armijas laikiem – **Armands Zeihmanis**."*

(no tvnet.lv)

- ***Bojāra*** → Gundars Bojārs, dz. 1967

  https://lv.wikipedia.org/wiki/Gundars_Bojārs

- **Otra figūra** = **Bojāra** = **eksmēra**

- ***Bojārs*** ↔ ***Zeihmanis***: draugs, padomnieks

NEL: General Approach

- Named Entity Recognition
- Candidate Selection

  Selecting possible **candidates** from the target *Knowledge Base*

- Disambiguation

  Deciding which candidate is the **correct identity** corresponding to the mention of a Named Entity

NEL: Context Free Approach

- Extract surface forms from KB or annotated corpus
  - *DBpedia* labels (rather sparse)
  - Internal links of *Wikipedia*
- Clean and catalog
- Fast string match

+ Simple
+ High precision
- Low recall
- Does not solve ambiguities

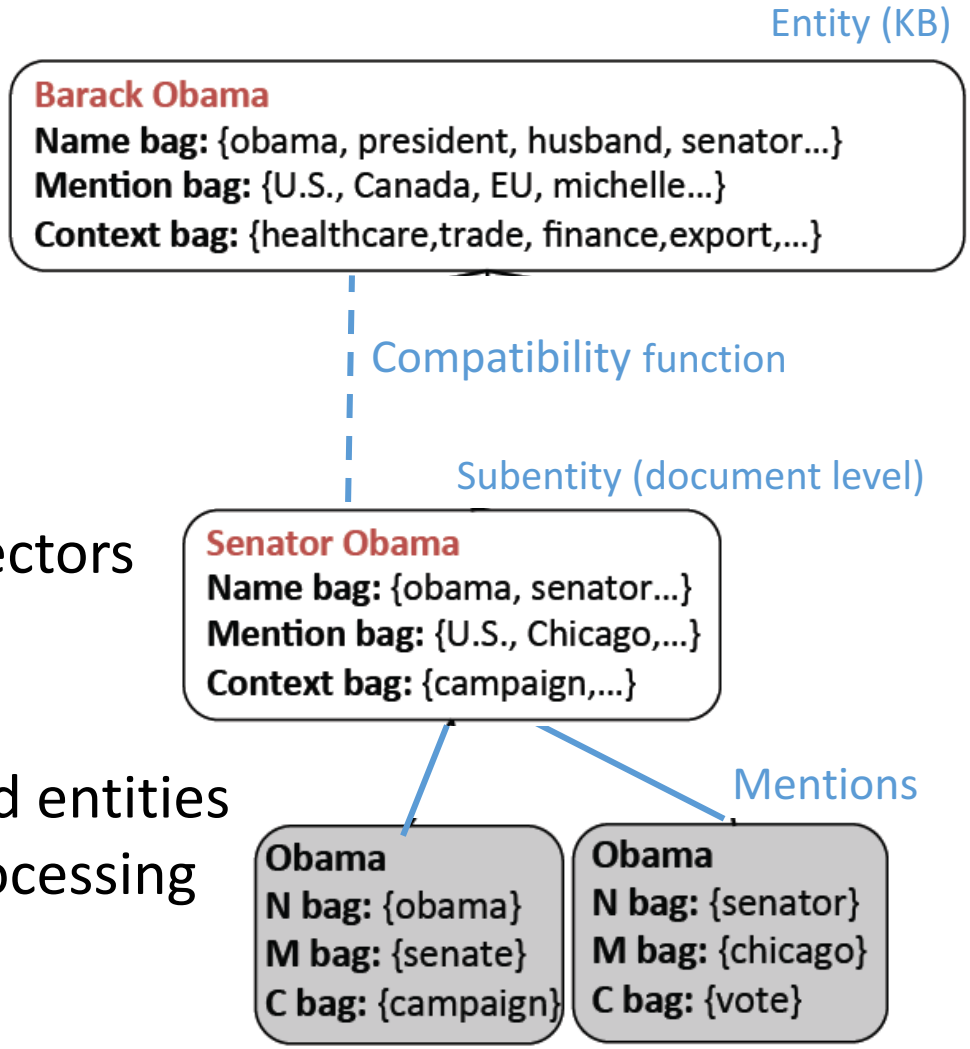
Generate name alternatives

Decide on which surface forms have ambiguous labels which cannot be considered without context

$$U_{l_i} = \left\{ e_i \in E_{l_i} : C(e_i) \geq \alpha \times \sum_{k}^{|E_{l_i}|} C(e_k) \right\}$$

NEL: Knowledge Rich Approach

- String match based candidate selection
- Mention attributes:
  - Bag-of-**name-words**
  - Bag-of-**mention-words**
  - Bag-of-**context-words**
- Cosine similarity between vectors



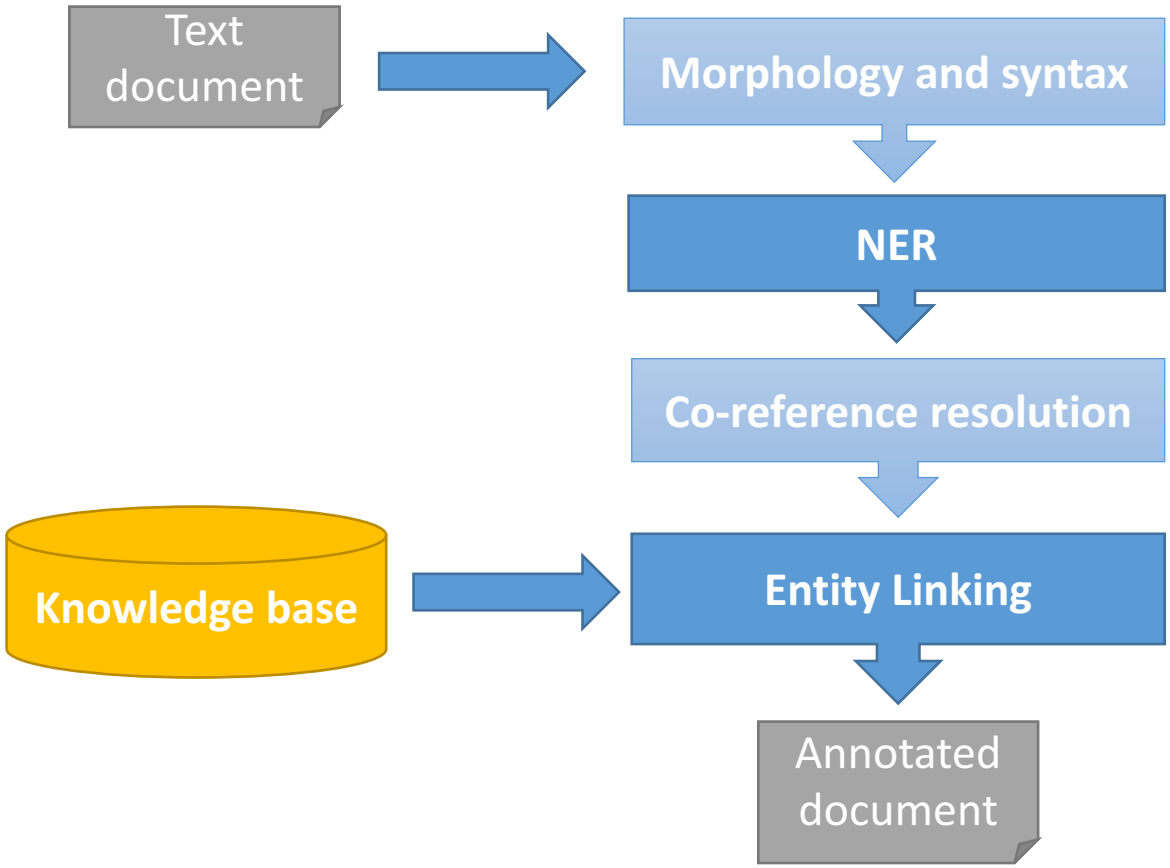
\+ Can solve ambiguities

\+ KB can be enriched with validated entities

\- Performance depends on pre-processing components

Wick et al, 2013; Stoyanov et al, 2012

NEL: Processing Pipeline

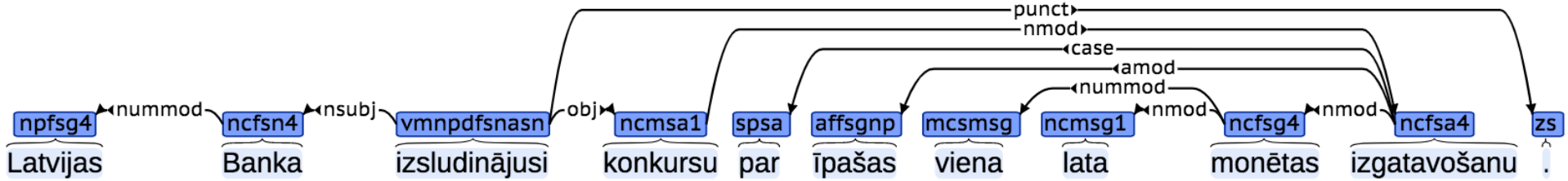


Morphology and Syntax

- Lexicon based morphological analyzer
- CMM morphological tagger
- LSTM transition-based parser
  - Word embeddings
  - Character LSTM representation

- Inflection generation
- Mention candidate selection

Model	UD (UAS, %)
LSTM + WE	75.1
LSTM + WE + morpho-tag	**75.4**



Paikens et al, 2013; Pretkalniņa et al, 2016

Named Entity Recognition

Model	NER (F1, %)
Baseline CRF	83.4
LSTM-CRF	63.7
LSTM-CRF + WE	**90.5**

GPE

Pagājušās nedēļas nogalē Valgā 32 kaimiņpilsētu šahisti cīnījās

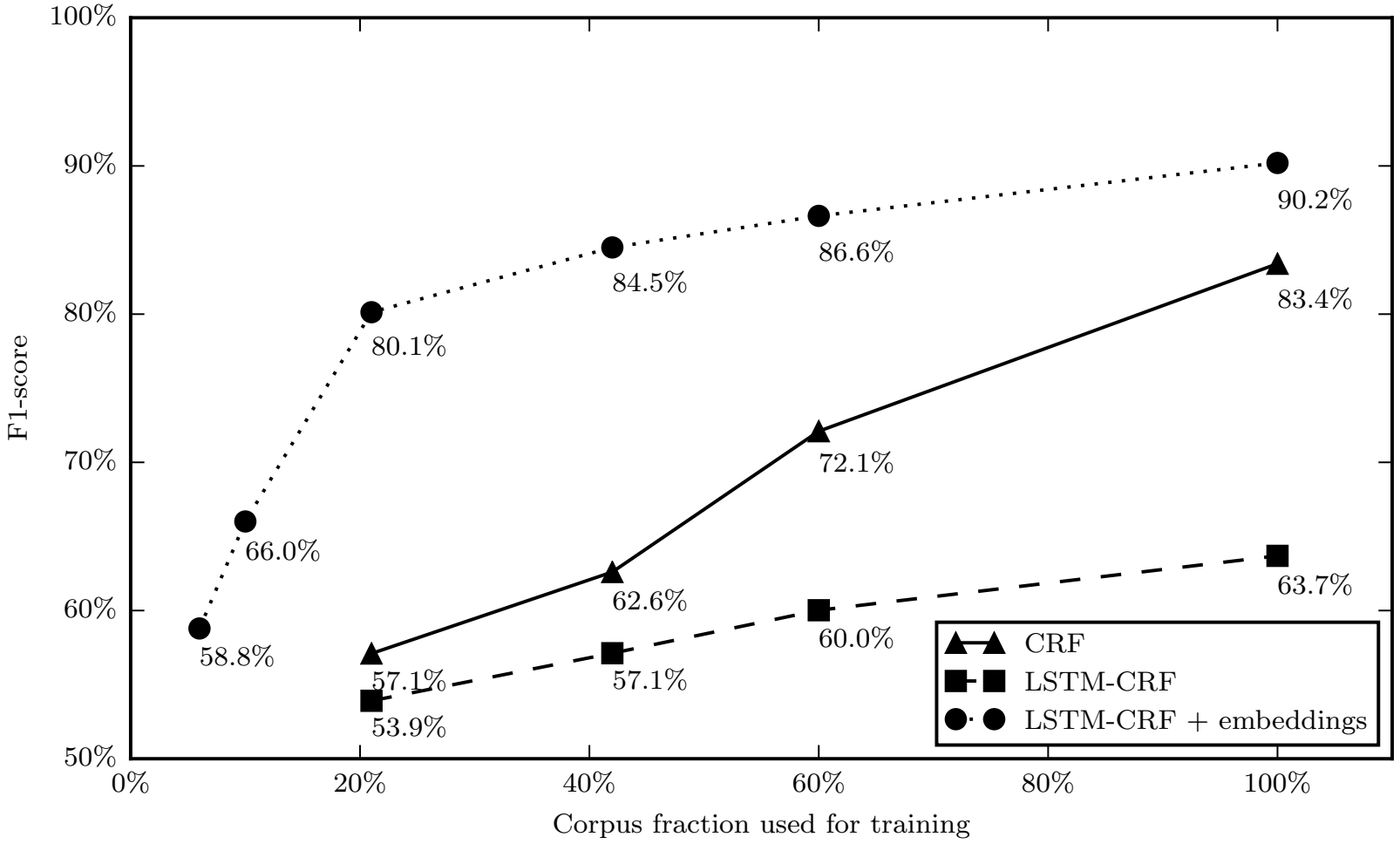
person event

Paula Keresa dzimšanas dienai veltītajā šaha turnīrā.

person

Atkal uzvarēja Māris Koops.

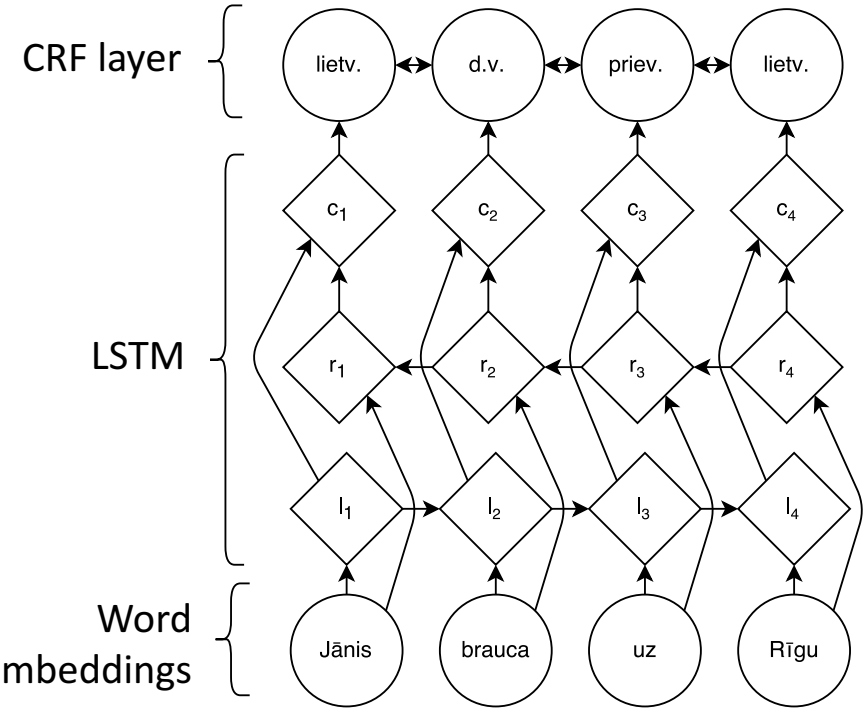
NER: Learning Curve

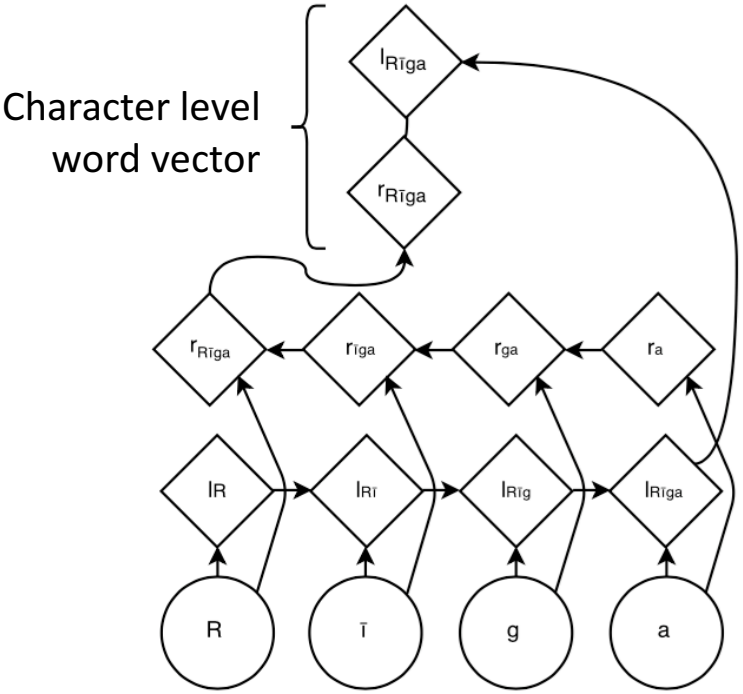


Named Entity recognition

Bi-directional LSTM-CRF

- Pre-trained word embeddings
- Character LSTM or CNN representations
- Joint cross-label modelling





Lample et al, 2015

Co-reference resolution

- Mentions that refers to the same real word entity

*Following in the footsteps of **their father Dainis**, **an ex-bobsleigh specialist**, **brothers Martins and Tomass Dukurs** are currently ranked among the globe's top skeleton racers. **Martins**, **world number one**, is **the gold medal favourite. He** dreams of sharing the podium with **his elder sibling**.*

- { their father Dainis, an ex-bobsleigh specialist }
- { their, brothers Martins and Tomass Dukurs }
- { Martins, Martins, world number one, the gold medal favourite, He, his }
- { Tomass Dukurs, his elder sibling }

Co-reference Resolution

- Mention detection

  NE, pronouns, noun phrases, etc.

- Mention chaining – clustering
  - Grammatically related mentions
  - Salience (gender, number, person, etc.)
  - Pairwise or cluster based
- Representative name and type (person, etc.)

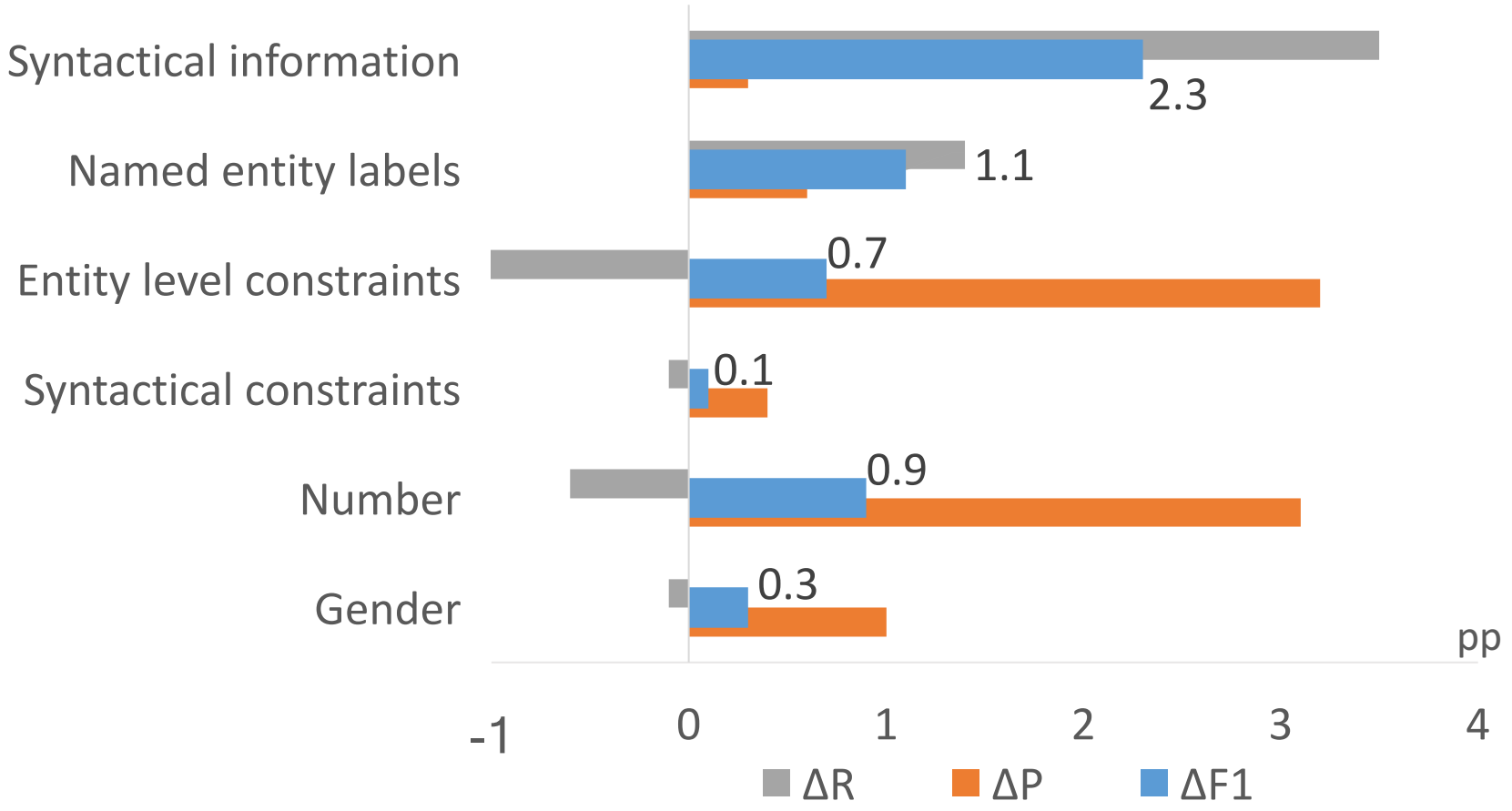
LVCoref: Rule Based System

- Exact string match
- Precise grammatical constructions
  - Appositives *(Andris Bērziņš, the president of Latvia)*
  - Nominal predicatives *(Jānis Bērziņš is a professor)*
  - Acronyms
- Head match variations *(Supreme Court of the Republic of Latvia ↔ Supreme Court of Latvia)*
- Pronoun anaphora

LVCoref: Results

	F1	P	R
Predicted mentions			
Exact match	40.0	74.2	28.0
+ Precise construction	46.2	74.5	34.3
+ Head match	56.2	69.4	47.3
+ Pronouns	58.0	65.1	52.3
Gold mentions			
Exact match	42.8	86.0	29.2
+ Precise construction	45.4	98.4	29.5
+ Head match	66.8	88.5	54.1
+ Pronouns	76.5	87.0	68.9

LVCoref: Components



Data annotation for Latvian

- Selected paragraphs from The Balanced Corpus of Modern Latvian
- NER/NEL
  - 8 categories: person, organization, location, GPE, product, event, time, entity
  - Derived from MUC and AMR guidelines
  - Links to Wikipedia
- Co-references:
  - Named entities, noun phrases and pronouns referring to specific entities
  - Derived from OntoNotes 6.0 guidelines
- Also syntax, frames, AMR

Projects:

- *Teksta automātiskas datorlingvistikas analīzes pētījums jauna informācijas arhīva produkta izstrādē (2013-2014)*
- *Daudzslāņu valodas resursu kopa teksta semantiskajai analīzei un sintēzei latviešu valodā (2016-2019)*

Future Work

- More annotated data
- Improve mention detection for Latvian
  - High impact on CR
  - Joint learning with shallow syntax parsing
- Named entity recognition
  - Hierarchical named entity mentions
  - Incorporate neural LSTM character level language model
- Co-reference resolution

  Joint learning with NEL
- Named entity recognition in speech

  Subword and class based language models
- Cross-lingual named entity linking

Publikācijas

- Miranda, S., Znotins, A., Cohen, S.B., Barzdins, G. (2017). Multilingual Clustering of Streaming News. Submitted to NAACL HLT.
- Znotiņš, A. (2016). Word Embeddings for Latvian Natural Language Processing Tools. In *Human Language Technologies: The Baltic Perspective*, IOS Press. Web of Science.
- Znotiņš, A., Polis, K., and Darģis R. (2015). Media monitoring system for Latvian radio and TV broadcasts. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 732–733. Web of Science, Scopus.
- Darģis, R. and Znotiņš, A. (2014). Baseline for Keyword Spotting in Latvian Broadcast Speech. In *Human Language Technologies: The Baltic Perspective*, IOS Press, 75–82. Web of Science.
- Znotiņš, A. and Paikens, P. (2014). Coreference Resolution for Latvian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, 3209–3213. Web of Science.
- Pretkalniņa, L., Znotiņš, A., Rituma, L., Goško, D. (2014). Dependency parsing representation effects on the accuracy of semantic applications — an example of an inflective language. In Proc*eedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, 4074–4081. Web of Science.

Summary

- **NER/NEL**: find all entity mentions within a text and link them to authoritative data source (knowledge base)
- NEL as text anchoring:
  - Coarse summary of what text is about
  - Other tasks can utilize information from KB
  - Other tasks can enrich named entities with additional semantic annotation layers