Modernising historical words

Tomaž Erjavec[1] Yves Scherrer[2]

[1]Dept. of Knowledge Technologies, Jožef Stefan Institute
Slovenia

[2]LATL-CUI, Université de Genève
Switzerland

Workshop on Exploring Historical Sources with Language Technology:
Results and Perspectives
December 2014, Den Haag

Outline

Variability of historical forms

word	lc	lemma	Freq
ljubesni	ljubezni	ljubezen	199
ljubesin	ljubezen	ljubezen	113
lubeſn	ljubezen	ljubezen	81
lubęsni	ljubezni	ljubezen	71
lubesni	ljubezni	ljubezen	67
lubęsen	ljubezen	ljubezen	65
lubeſne	ljubezni	ljubezen	63
lubesen	ljubezen	ljubezen	59
ljubezni	ljubezni	ljubezen	58
lubeſni	ljubezni	ljubezen	50
ljubezen	ljubezen	ljubezen	46
ljubesnijo	ljubeznijo	ljubezen	39
ljubesen	ljubezen	ljubezen	39
ljubésni	ljubezni	ljubezen	23
Lubeſn	ljubezen	ljubezen	18
lubesn	ljubezen	ljubezen	15
ljubésen	ljubezen	ljubezen	15
ljubezin	ljubezen	ljubezen	14
Lubesni	ljubezni	ljubezen	13
lubesne	ljubezni	ljubezen	12
Lubęsen	ljubezen	ljubezen	11
lubęsnio	ljubeznijo	ljubezen	8
Ljubesin	ljubezen	ljubezen	8
lubésen	ljubezen	ljubezen	7
lubesnio	ljubeznijo	ljubezen	7
Lubesen	ljubezen	ljubezen	6
Ljubezni	ljubezni	ljubezen	6
Ljubezen	ljubezen	ljubezen	5
ljubézni	ljubézni	ljubezen	4
ljubeznijo	ljubeznijo	ljubezen	4
Ljubésni	ljubezni	ljubezen	4
lubiesen	ljubezen	ljubezen	3
ljubésin	ljubezen	ljubezen	3
lubęsèn	ljubezen	ljubezen	2

Motivation

Why modernise historical words:

- Linguistic annotation:
  Automatic PoS and lemma annotation can be performed with models for contemporary language
- Information retrieval:
  Enables search in cultural heritage digital libraries and corpora by modern word (lemma)
- Comprehension:
  Easier to read old texts with modernised words

Outline

Slovene historical language

- Part of Austro(-Hungarian) empire till 1918; dominant written language was German
- Change of alphabet ∼1840: Bohorič (long s + digraphs, e.g. zh) to Gaj (c,s,z, č,š,ž)
- Slow to standardise orthography
- Many very different dialects, reflected in the spelling

1790	Al	ta	nar bòl	vashna	reſsniza	je	moja	lubęsen	prut	Nęshki.
(18B)	*ali*	*ta*	*najbolj*	*važna*	*resnica*	*je*	*moja*	*ljubezen*	*proti*	*nežki*

1843	poboshnim	ſerzam	in	veſtjo	pridnoſt	in	ljubesin	k	ſvojimu	ſtanu	sdrushi
(19A)	*pobožnim*	*srcem*	*in*	*vestjo*	*pridnost*	*in*	*ljubezen*	*k*	*svojemu*	*stanu*	*združi*

1872	Otroška	ljubezen	naj	zmír	te	navdaja	Za	starše,	za	brate,	Bogá	in	cesarja
(19B)	*otroška*	*ljubezen*	*naj*	*zmeraj*	*te*	*navdaja*	*za*	*starše,*	*za*	*brate,*	*boga*	*in*	*cesarja*

IMP resources

Overview:

- Result of several projects (AHLib, EU IMPACT, Google award)
- A BLARK for historical Slovene
  - 1584–1919, most texts from > 1850
  - digital library (658 units, 46,645 pages)
  - lexicon (21,653 lem., 51,156 contemp. & 73,263 histo.)
  - hand annotated corpus (267,124 words)
  - annotation toolchain (DL → corpus 14,358,423 words)
- For HLT: XML TEI & CC BY
- For DH: HTML & noSketchEngine
- http://nl.ijs.si/imp/

IMP resources

Overview:

- Result of several projects (AHLib, EU IMPACT, Google award)
- A BLARK for historical Slovene
  - 1584–1919, most texts from > 1850
  - digital library (658 units, 46,645 pages)
  - lexicon (21,653 lem., 51,156 contemp. & 73,263 histo.)
  - hand annotated corpus (267,124 words)
  - annotation toolchain (DL $\rightarrow$ corpus 14,358,423 words)
- For HLT: XML TEI & CC BY
- For DH: HTML & noSketchEngine
- http://nl.ijs.si/imp/

Annotation toolchain

ToTrTaLe (Erjavec, 2011):

- Tokenises and sentence segments the text
- Transcribes the words to contemporary spelling
- PoS (MSD) tags the contemporary words
- Lemmatises the PoS tagged contemporary words
- TEI P5 I/O

Transcription:

- Uses hand written rules (e.g. cov$ $\rightarrow$ cev$ for stricov $\rightarrow$ stricev)
- Vaam applies all the rules to a word and produces a set of results
- These are filtered against a lexicon of contemporary word forms
- As the result take the most frequent word form

Annotation toolchain

ToTrTaLe (Erjavec, 2011):

- Tokenises and sentence segments the text
- Transcribes the words to contemporary spelling
- PoS (MSD) tags the contemporary words
- Lemmatises the PoS tagged contemporary words
- TEI P5 I/O

Transcription:

- Uses hand written rules (e.g. cov$ $\rightarrow$ cev$ for stricov $\rightarrow$ stricev)
- Vaam applies all the rules to a word and produces a set of results
- These are filtered against a lexicon of contemporary word forms
- As the result take the most frequent word form

New approach to transcription

- Problems with ToTrTaLe transcription:
  - Problem with low coverage ($\sim$100 rules not enough)
  - Experiment showed low precision ($\sim$72% on OOV words)
- IMP lexicon:
  - Available dataset with $\langle historicalword, contemporaryword \rangle$ pairs
  - Can we automatically train a transcription system?

Outline

Character-based MT for modernisation

Hypothesis:

- Historical and contemporary language words may be viewed as closely related language varieties
- So we can use machine translation to transcribe between them, taking a character as a “word”

Word-level SMT:						Character-level SMT:						
EN	I	go	to	Paris	.	SL-old	s	o	l	n	c	e
	\	/	|	|	|		|	|	\	/	|	|
SL	Grem		v	Pariz	.	SL	s	o		n	c	e

Background

- The statistical translation model can be trained on the lexicon
- Not the first / only ones to think of this:
  - (Vilar et al. 2007; Tiedemann 2009)
  - (Sánchez-Martínez et al. 2013; Pettersson et al. 2013)
- We use Moses STM for our experiments
- Reported on this experiment in:
  Scherrer & Erjavec: Modernizing historical Slovene words with character-based SMT. Proceedings of the 4th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013), ACL.

Two experiments

- Supervised:
  - Make use of manually annotated $\langle historical, contemporary \rangle$ word pairs
- Unsupervised:
  - Use "monolingual" data only: $\langle historical \rangle + \langle contemporary \rangle$

The dataset

- Lexicons extracted from manually annotated corpora, in three 50-year slices:
  - 1750 – 1800 [18B]
  - 1800 – 1850 [19A]
  - 1850 – 1900 [19B]
- A lexicon of contemporary Slovene

Normalised historical form

- convert Bohorič to Gaj alphabet (with rules)
- lower-case
- remove vowel diacritics

Historical Slovene lexicons

L*goo*

- Lexicon extracted from fully annotated goo300k corpus
- Normalised historical form, modernised form, frequency per time period
- 18B: 6,000 entries, 19A: 18,000 entries, 19B: 30,000 entries
- Serves as training set

L*foo*

- Lexicon extracted from partially annotated foo3M corpus
- Words disjoint from L*goo*
- Serves as a realistic test set

Historical Slovene lexicons

L*goo*

- Lexicon extracted from fully annotated goo300k corpus
- Normalised historical form, modernised form, frequency per time period
- 18B: 6,000 entries, 19A: 18,000 entries, 19B: 30,000 entries
- Serves as training set

L*foo*

- Lexicon extracted from partially annotated foo3M corpus
- Words disjoint from L*goo*
- Serves as a realistic test set

Example entries

bčelnemu	čebelnemu	19B:1
bdenjam	bedenjem	19A:1
bdi	bedi	19A:1
bdijo	bedijo	19A:1
bebasta	bebasta	19B:1
bebca	bebca	19B:1
be	bi	18B:35
beda	beda	19B:1
bega	bega	19A:1
begam	begom	19A:1
begate	begate	19A:1
begati	begati	19B:1
beg	beg	19A:2 19B:3
begu	begu	19A:2 19B:2

Contemporary Slovene lexicon

Sloleks

- Word forms annotated with lemmas, MSD tags, frequency (number of occurrences in Gigafida reference corpus)
- 930k lower-cased word forms (100k lemmas)
- Result of SSJ project, www.slovenscina.eu (CC BY-NC)

Outline

Supervised experiment

Goal:

Automatically modernise historical Slovene words
using character-based statistical machine translation (CSMT)

- Train a CSMT model with ⟨*historical*, *contemporary*⟩ word pairs from the L*goo* lexicon
  - Tools: GIZA++, Moses, IRSTLM
  - 5-gram (character) language model trained on Sloleks
  - No distortion (swap operations)
  - MERT on 20% of the training data
- One model per time period

→ Infer regularities in character correspondences

Supervised experiment

Goal:

Automatically modernise historical Slovene words
using character-based statistical machine translation (CSMT)

- Train a CSMT model with ⟨*historical*, *contemporary*⟩ word pairs from the L*goo* lexicon
  - Tools: GIZA++, Moses, IRSTLM
  - 5-gram (character) language model trained on Sloleks
  - No distortion (swap operations)
  - MERT on 20% of the training data
- One model per time period

→ Infer regularities in character correspondences

Unsupervised experiment

Goal:
Automatically modernise historical Slovene words
using character-based statistical machine translation (CSMT)

- Do not use the manual annotations of L*goo*
- Create a noisy list of ⟨*historical* , *contemporary*⟩ word pairs
  - Historical words from L*goo*
  - Contemporary words from Sloleks
- Train a CSMT model with these noisy word pairs
  - Same parameters, but no MERT
- One model per time period

→ Infer regularities in character correspondences
→ Eliminate some of the noise in the training data

Unsupervised experiment

Goal:
Automatically modernise historical Slovene words
using character-based statistical machine translation (CSMT)

- Do not use the manual annotations of L*goo*
- Create a noisy list of $\langle$*historical*, *contemporary*$\rangle$ word pairs
  - Historical words from L*goo*
  - Contemporary words from Sloleks
- Train a CSMT model with these noisy word pairs
  - Same parameters, but no MERT
- One model per time period

$\rightarrow$ Infer regularities in character correspondences
$\rightarrow$ Eliminate some of the noise in the training data

Unsupervised experiment

How to create the noisy word pairs?

BI-SIM

A measure of formal similarity based on bigrams (Kondrak & Dorr 2004)

- Convert strings to bigram sequences
- Count identical bigrams (+1) and semi-identical bigrams (+0.5)
- Normalise by the length of the longer string
- 1 $\rightarrow$ the words are identical, 0 $\rightarrow$ no character matches

- For each historical word [from L*goo*], choose the modern word(s) [from Sloleks] with the highest BI-SIM value
- Discard word pairs with BI-SIM value lower than 0.8 (empirically chosen threshold)

Unsupervised experiment

How to create the noisy word pairs?

BI-SIM

A measure of formal similarity based on bigrams (Kondrak & Dorr 2004)

- Convert strings to bigram sequences
- Count identical bigrams (+1) and semi-identical bigrams (+0.5)
- Normalise by the length of the longer string
- 1 $\rightarrow$ the words are identical, 0 $\rightarrow$ no character matches

- For each historical word [from L*goo*], choose the modern word(s) [from Sloleks] with the highest BI-SIM value
- Discard word pairs with BI-SIM value lower than 0.8 (empirically chosen threshold)

Unsupervised experiment

How to create the noisy word pairs?

BI-SIM

A measure of formal similarity based on bigrams (Kondrak & Dorr 2004)

- Convert strings to bigram sequences
- Count identical bigrams (+1) and semi-identical bigrams (+0.5)
- Normalise by the length of the longer string
- 1 $\rightarrow$ the words are identical, 0 $\rightarrow$ no character matches

- For each historical word [from L*goo*], choose the modern word(s) [from Sloleks] with the highest BI-SIM value
- Discard word pairs with BI-SIM value lower than 0.8 (empirically chosen threshold)

Outline

Results

- 3 time periods
- 2 experiments (supervised and unsupervised)
- With and without lexicon filter

Lexicon filter

- The candidates proposed by the CSMT system are not necessarily existing modern Slovene words.
- Without lexicon filter: take CSMT candidate with highest score
- With lexicon filter: take CSMT candidate with highest score that occurs in Sloleks

- Baseline: Identical word pairs

Results

- 3 time periods
- 2 experiments (supervised and unsupervised)
- With and without lexicon filter

Lexicon filter

- The candidates proposed by the CSMT system are not necessarily existing modern Slovene words.
- Without lexicon filter: take CSMT candidate with highest score
- With lexicon filter: take CSMT candidate with highest score that occurs in Sloleks

- Baseline: Identical word pairs

Results

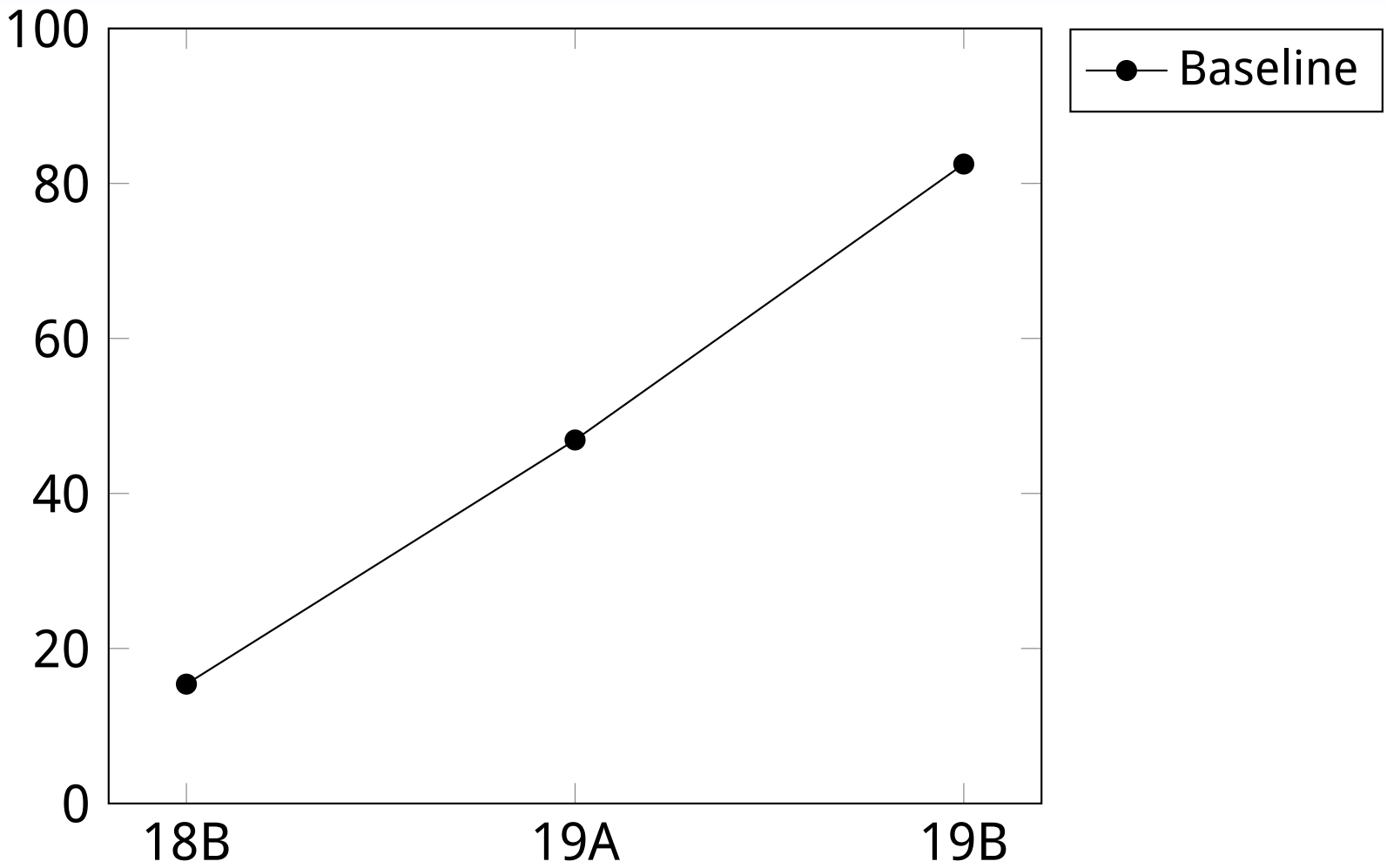
- 3 time periods
- 2 experiments (supervised and unsupervised)
- With and without lexicon filter

Lexicon filter

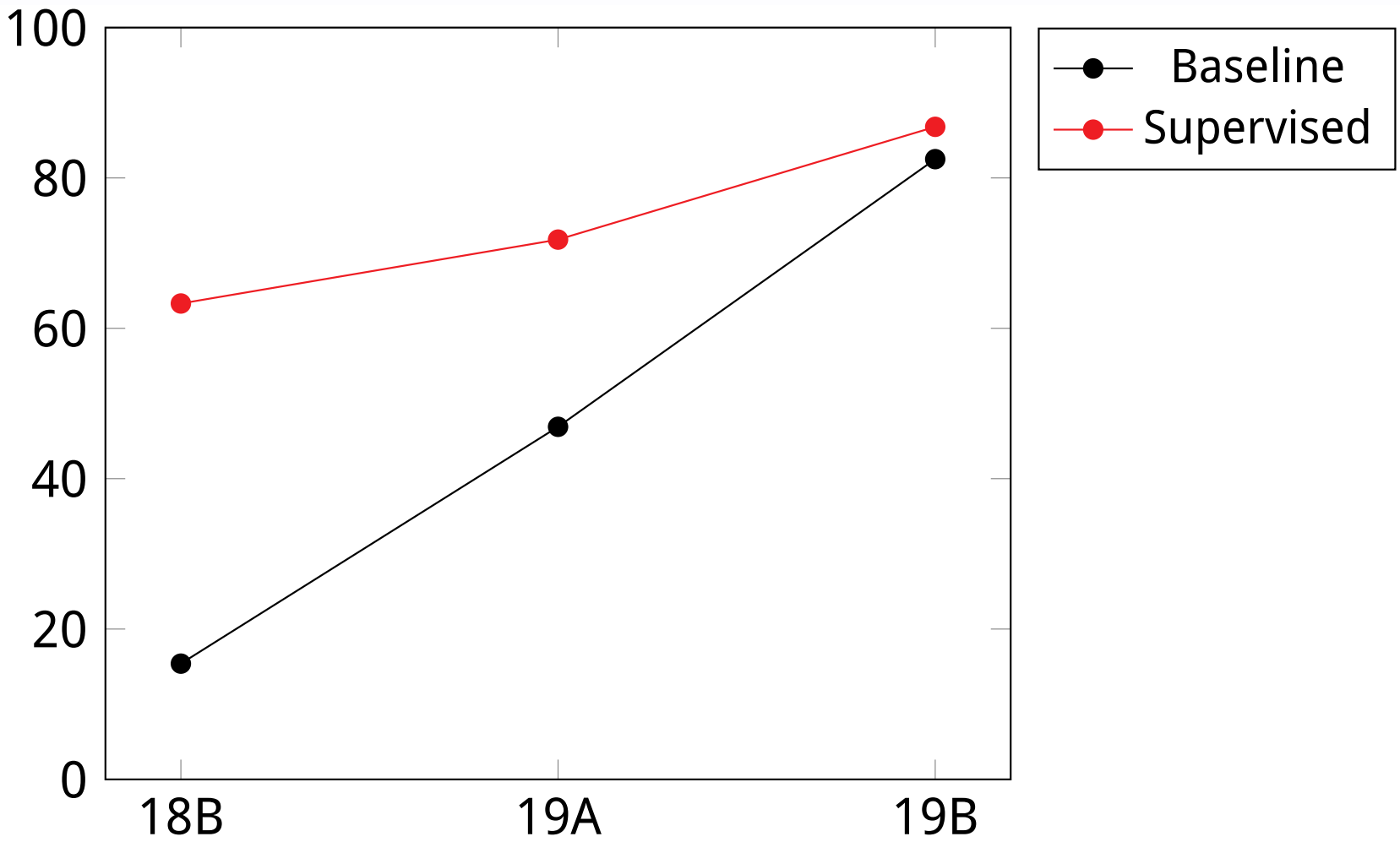
- The candidates proposed by the CSMT system are not necessarily existing modern Slovene words.
- Without lexicon filter: take CSMT candidate with highest score
- With lexicon filter: take CSMT candidate with highest score that occurs in Sloleks

- Baseline: Identical word pairs

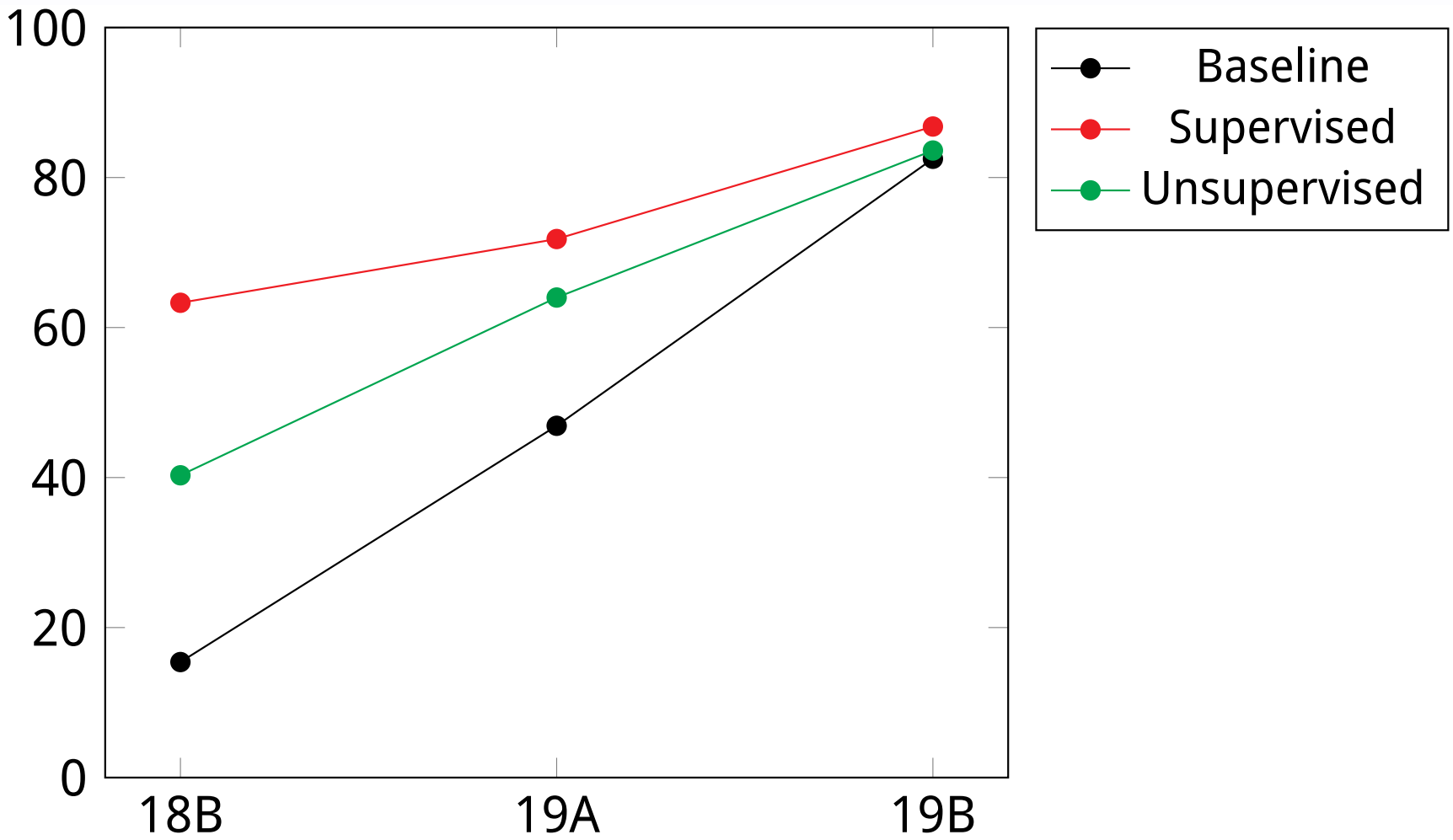
Results



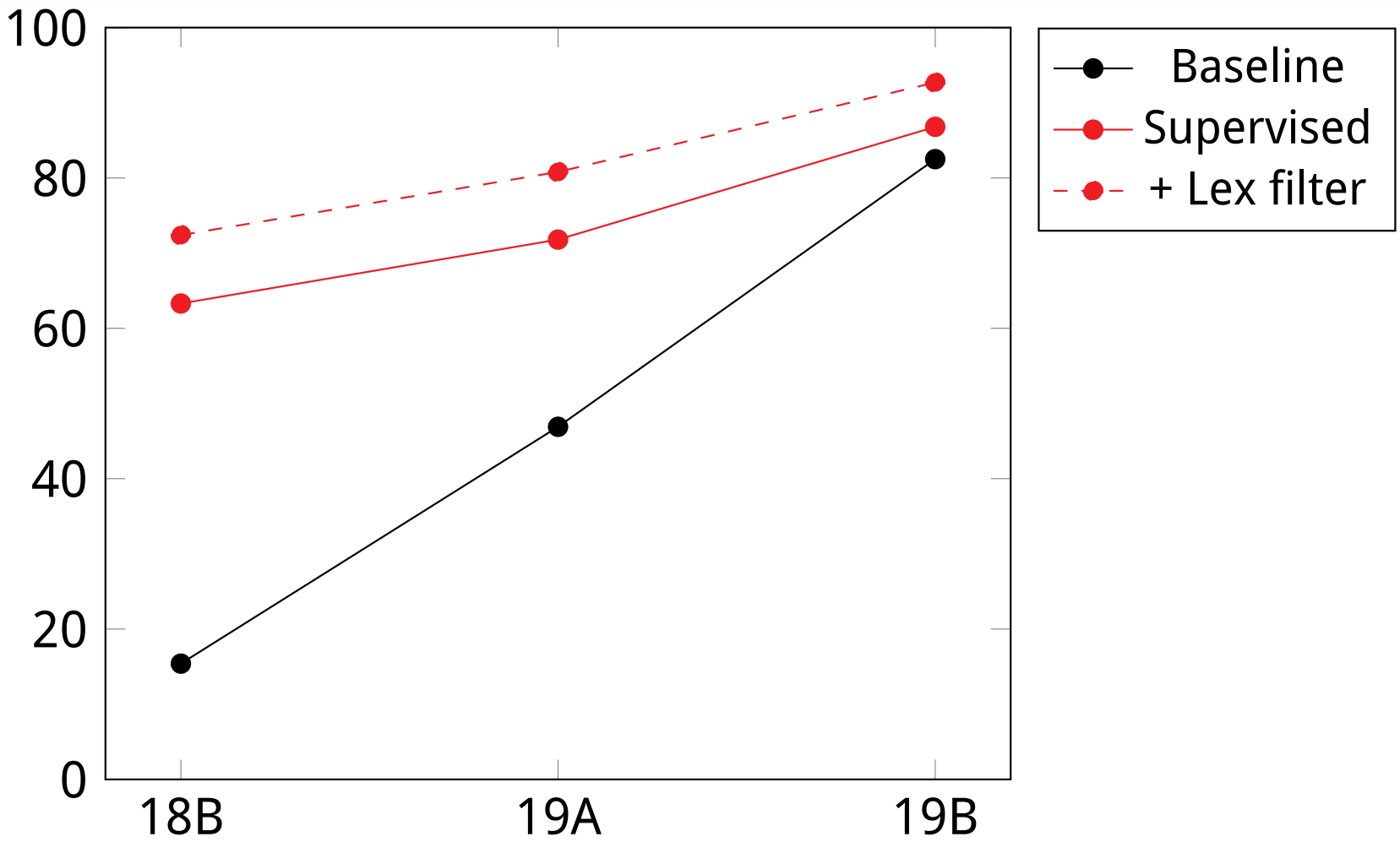
Results



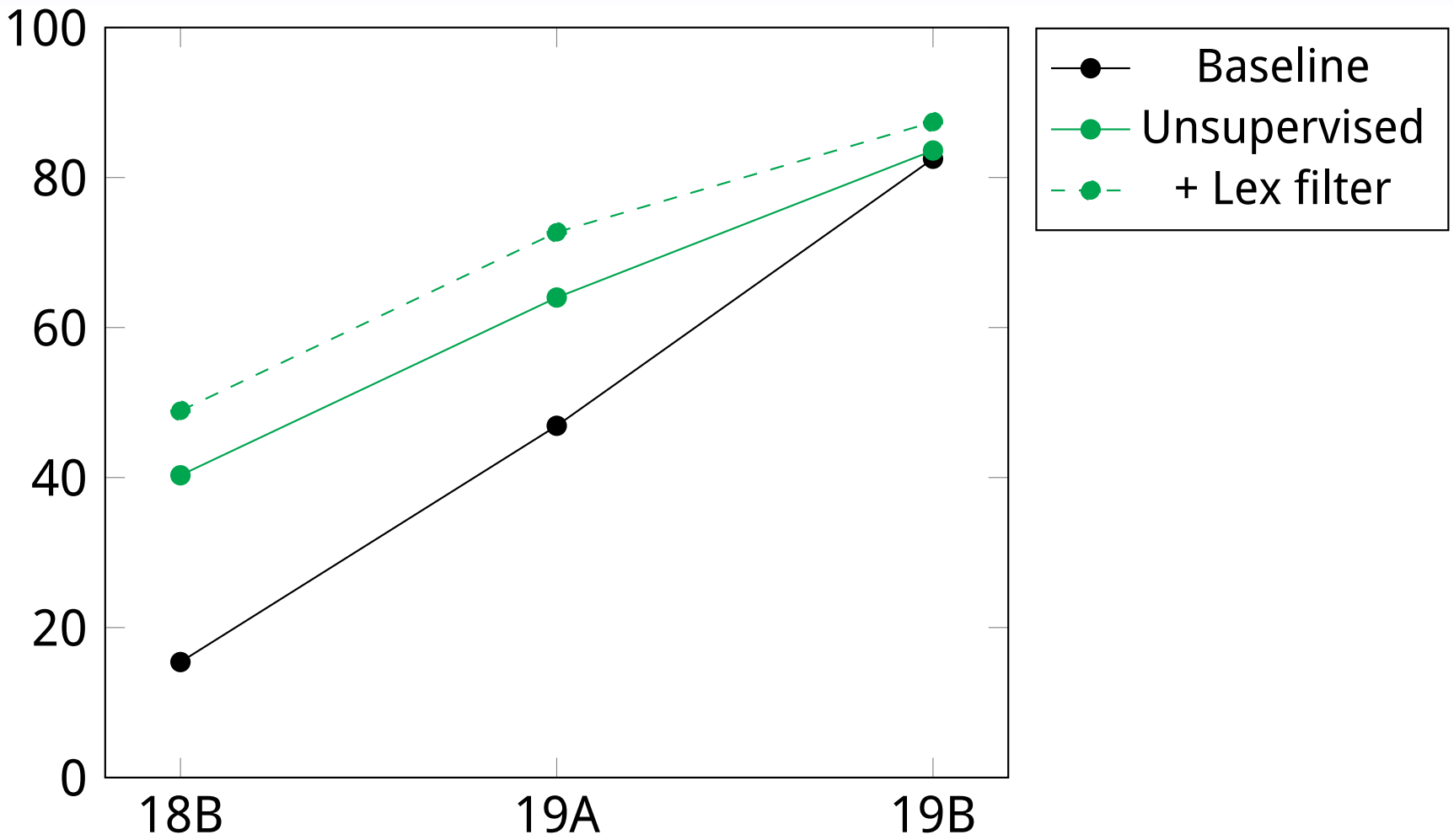
Results


100
80
60
40
20
0
18B
19A
19B
Baseline
Supervised
Unsupervised

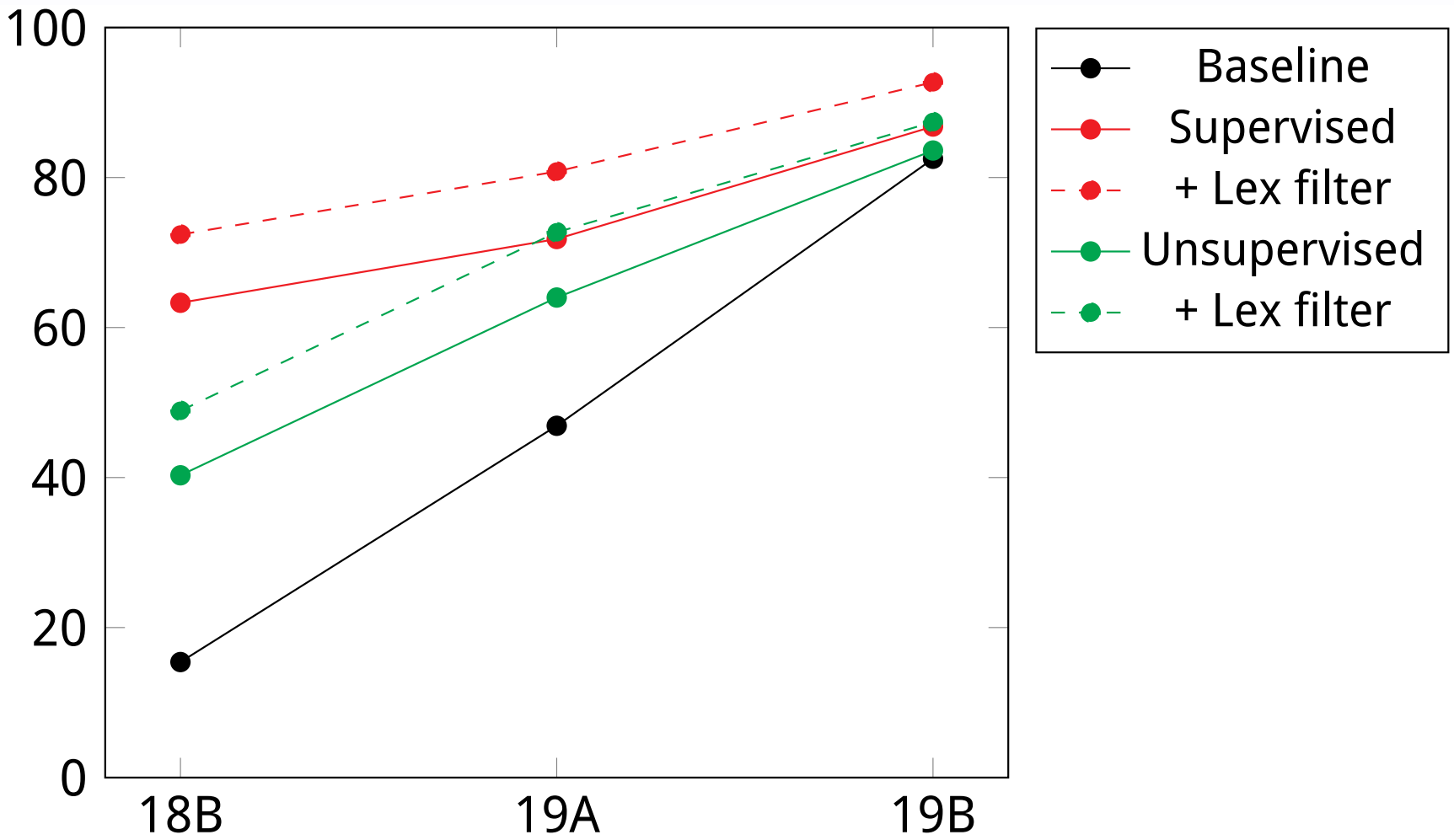
Results



Results


100
80
60
40
20
0
18B
19A
19B
Baseline
Unsupervised
+ Lex filter

Results


100
80
60
40
20
0
18B
19A
19B
Baseline
Supervised
+ Lex filter
Unsupervised
+ Lex filter

Outline

Conclusion

- Supervised experiment: +57.0% absolute on baseline (18B)
  - Simulates the task of annotating new texts of a known language and period
- Unsupervised experiment: +33.5% absolute on baseline (18B)
  - Simulates the task of annotating texts of an unknown language or period
- All experiments beat the baseline, even on the difficult 19B set

- Tried similar method: Ljubešić, N. Erjavec, T., Fišer, D. Standardizing tweets with character-level machine translation. Proc. of CICLing 2014.

Conclusion

- Supervised experiment: +57.0% absolute on baseline (18B)
  - Simulates the task of annotating new texts of a known language and period
- Unsupervised experiment: +33.5% absolute on baseline (18B)
  - Simulates the task of annotating texts of an unknown language or period
- All experiments beat the baseline, even on the difficult 19B set
- Tried similar method: Ljubešić, N. Erjavec, T., Fišer, D. Standardizing tweets with character-level machine translation. Proc. of CICLing 2014.

Conclusion

Future work:

- Handle tokenization changes
  - 1 word $\leftrightarrow$ 2 words (e.g. nar bolj $\rightarrow$ najbolj)
- Move from a deterministic setting
- Effect of background resources (contemporary lexicon)
- Application (precision v.s. recall)