Leveraging Inflection Tables for Stemming and Lemmatization

Garrett Nicolai and Greg Kondrak

Department of Computing Science

University of Alberta

1. Inflectional Simplification

- Many languages contain many surface forms of a single dictionary word.
- This results in very sparse data sets, with a large number of forms only appearing infrequently, even for very large corpora.
- Inflectional simplification aims to reduce this sparsity by simplifying all forms down to a small number of representative forms.
- Stemming and lemmatization are two common methods of reducing morphological complexity.

2. Stemming

What is Stemming?

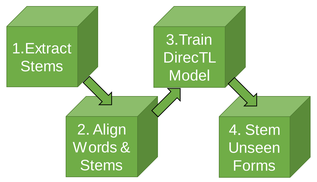
- Stemming is the removal of inflectional affixes from a word.
- For example, the stem of *playing* is *play*.
- It is one method of morphological simplification: reducing the type-to-token ratio.
- One related group of words may have several stems, and they do not need to be free morphemes.
- For example, *fly*, *flying*, *flew*, *flies,* etc. Has at least 3 different stems: *fly, fli,* and *flew.*
- Stems are abstract representations, and thus there is no consensus on how words *should* be stemmed.

	Singular 1st	2nd	3rd	Plural 1st
Present	*doy*	*das*	*da*	*damos*
Imperfect	*daba*	*dabas*	*daba*	*dábamos*
Preterite	*di*	*diste*	*dio*	*dimos*
Future	*daré*	*darás*	*dará*	*daramos*

A partial inflection table of the Spanish verb *dar.*

Word Form	Meaning	Tag	Stem
geben	*"to give"*	*INF*	*geb*
gibt	*"gives"*	*3SIE*	*gib*
gab	*"gave"*	*1SIA*	*gab*
gegeben	*"given"*	*PP*	*geb*

Basic Method (Word-to-Stem)



Unsupervised Stem Extraction

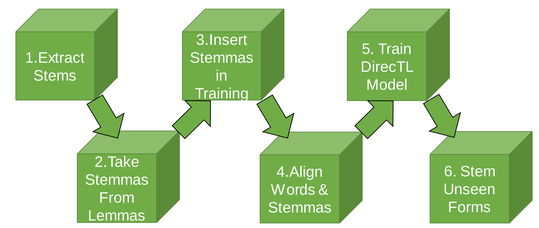
1. Inflection tables	
geben+2SIE	gibst
setzen+2SIE	setzt
PP+tun+PP	tust
PP+geben+PP	gegeben
PP+setzen+PP	gesetzt
tun+1SIA	getan

2. Abstract Tags	
STEM 2SIE	gibst
STEM 2SIE	setzt
STEM 2SIE	tust
PP STEM PP	gegeben
PP STEM PP	gesetzt
PP STEM PP	getan

3. Extract Stems	
STEM|2SIE|	gib|st|
STEM|2SIE|	setz|t|
STEM|2SIE|	tu|st|
PP|STEM|PP|	ge|geb|en|
PP|STEM|PP|	ge|setz|t|
PP|STEM|PP|	ge|ta|n|

- Stems *within* a table are fairly regular
- EM aligner maximizes joint likelihood of stems and affixes
- Affixes *across* tables are fairly regular
- Does not require expensive morphological annotation

Joint Method (Word-to-Stemma)



Transduction via DirecTL+

1. Align source / target pairs

Source	Target
g|i|b|t|	g|i|b|+t|
g|i|b|s|t|	|g|i|b|+s|t|
g|e|b|e|n|	g|e|b|+e|n|

2. Extract transduction rules

t ➔ +t / b_ s ➔ +s / gib_t

e ➔ +e / g:g e:e b:b _

3. Apply transduction rules

schreibst ➔ schreib+st

Example DirecTL+ Input / Output:

Word-to-Stem:

Input:

g|i|b|s|t| g|i|b|+s|t|

g|e|a|t|m|e|t| g|e+|a|t|m|+e|t|

s|c|h|r|e|i|b|e| s|c|h|r|e|i|b|+e|

Output:

gebe geb+e

schreibst schreib+st

Word-to-Stemma:

Input:

g|i|b|st| g|e|b|+2SIE|

ge|a|t|m|et| PP+|a|t|m|+PP|

s|c|h|r|e|i|b|e| s|c|h|r|e|i|b|+1SIE|

Output:

g|e|b|e| g|e|b|+1SIE|

s|c|h|r|e|i|b|st| s|c|h|r|e|i|b|+2SIE|

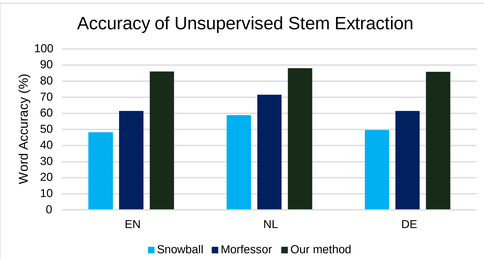
Specifics

- We evaluate our methods on 3 languages: English (EN), Dutch (NL), and German (DE).
- Training sets are extracted from either CELEX, or Wiktionary.
- Testing sets are extracted from CELEX.

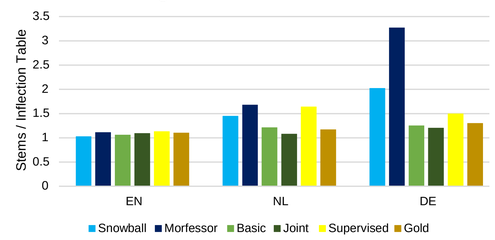
Interpretation

- Of the unsupervised methods, our method extracts the stems most consistent with human annotation.
- Our methods produce consistent stems.
- Our Joint method benefits from its access to morphological tags.
- Even when the training and testing data are from different sources, our methods produce high-quality stems.
- Our accuracy is similar to a system that require much more information.

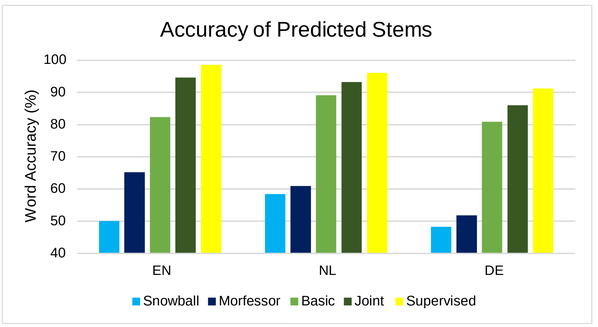
Accuracy of Unsupervised Stem Extraction

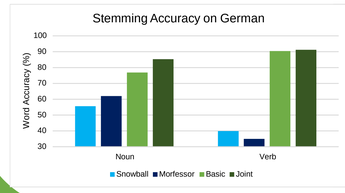


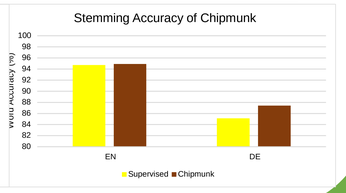
Average Stems Per Inflection Table



Accuracy of Predicted Stems







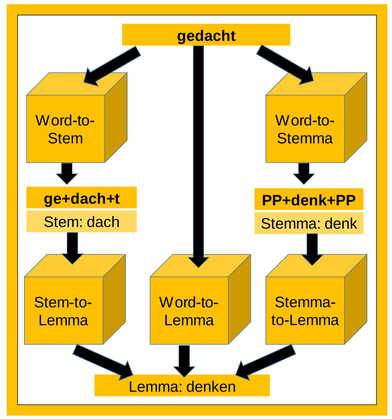
Conclusions

- The Joint method achieves almost supervised-level stems
- The supervised method approaches the state of the art.
- Our method reduces the number of different stems.
- The Joint method benefits from a type of POS awareness.
- In future, we hope to expand to languages with more complex morphology

3. Lemmatization

What is Lemmatization?

- Lemmatization is the process of restoring a word to its dictionary form, or *lemma*.
- Unlike stemming, the output of a lemmatizer must be a free morpheme.
- Likewise, there is only one correct lemma per word, but due to homonymity, the same surface form may lemmatize differently.
- We compose our word-to-*x* models with *x*-to-lemma models to recover lemmas: we first transduce a stem or stemma, and then a lemma.
- We perform context-free lemmatization.



Reranking

- We make use of an unannotated corpus to rerank an *n*-best list..
- Features include:
  - Normalized DirecTL score
  - Rank in the *n*-best list
  - Presence or absence in the corpus
  - Normalized likelihood from a 4-gram character language model

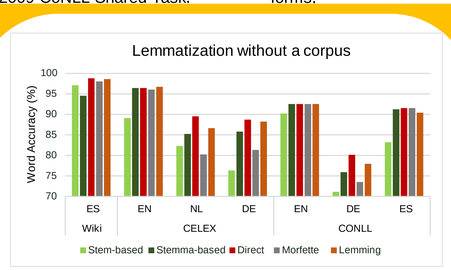
Intrinsic Evaluation

- In addition to the data sets we used for stemming, we introduce Spanish, extracted from Wiktionary, and English, German, and Spanish test sets from the 2009 CoNLL Shared Task.
- We evaluate against state-of-the-art Lemming, and Morfette,
- We consider a prediction correct if it matches a lemma of any of the potential surface forms.

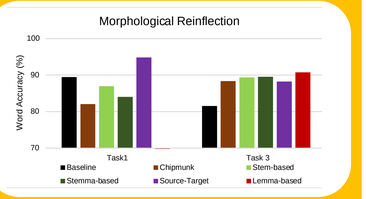
Extrinsic Evaluation

- We also perform an extrinsic evaluation using the data from the 2016 SIGMORPHON Shared Task on Morphological Reinflection.
- Task 1 predicts an inflected form from a lemma and morphological tag.
- Task 2 replaces the lemma of task 1 with an already inflected form.

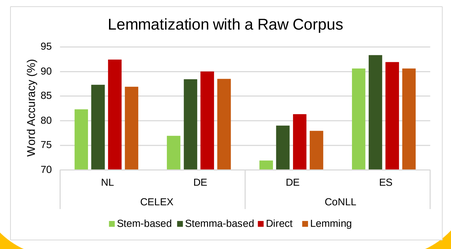
Lemmatization without a corpus



Morphological Reinflection



Lemmatization with a Raw Corpus



Conclusions

- For lemmatization, none of our composite methods are as accurate as our Direct model.
- Our Direct model is competitive with the state-of-the-art methods, and surpasses them when a corpus is introduced.
- For re-inflection, our lemma-based method out-performs all other methods, which is mirrored by the source-target model for Task 1.
- In the future, we hope to expand our methods to full morphological analysis.