Morphological Reinflection via Discriminative String Transduction

Garrett Nicolai **Bradley Hauer** **Adam St Arnaud** **Grzegorz Kondrak**

University of Alberta

1. Inflection and Re-Inflection

Morphological inflection
- Take a **lemma**, the basic "dictionary" form of a word
- Produce a particular **inflected form**
- Conveys additional information such as person (first person, third person, etc.), number (singular or plural), etc.
- **Reinflection** goes from one inflected form to another, starting with an inflected form rather than the lemma

2. Three Tasks

T1: Standard inflection:
take + "past participle" = *taken*

T2: Labelled reinflection:
Lemma(*taking,* "present participle") + "past participle" = *taken*

T3: Unlabelled reinflection:
Lemma(*taking*) + "past participle" = *taken*

We do not use the input tag in T2, instead treating T2 and T3 the same.

3. String Transduction

- We use DirecTL+, a feature-rich string transduction tool
- Learns models for transducing inflected forms from lemmas and tags, as in Nicolai et al. (2015)
- Tags are treated as atomic symbols
- For tasks 2 and 3, we also apply DirecTL+ to **lemmatize** the input forms, reducing T2 and T3 to T1.

4. Inflection and Reinflection as String Transduction

*Note: For T2, the T3 data is not used, per the rules of the task.
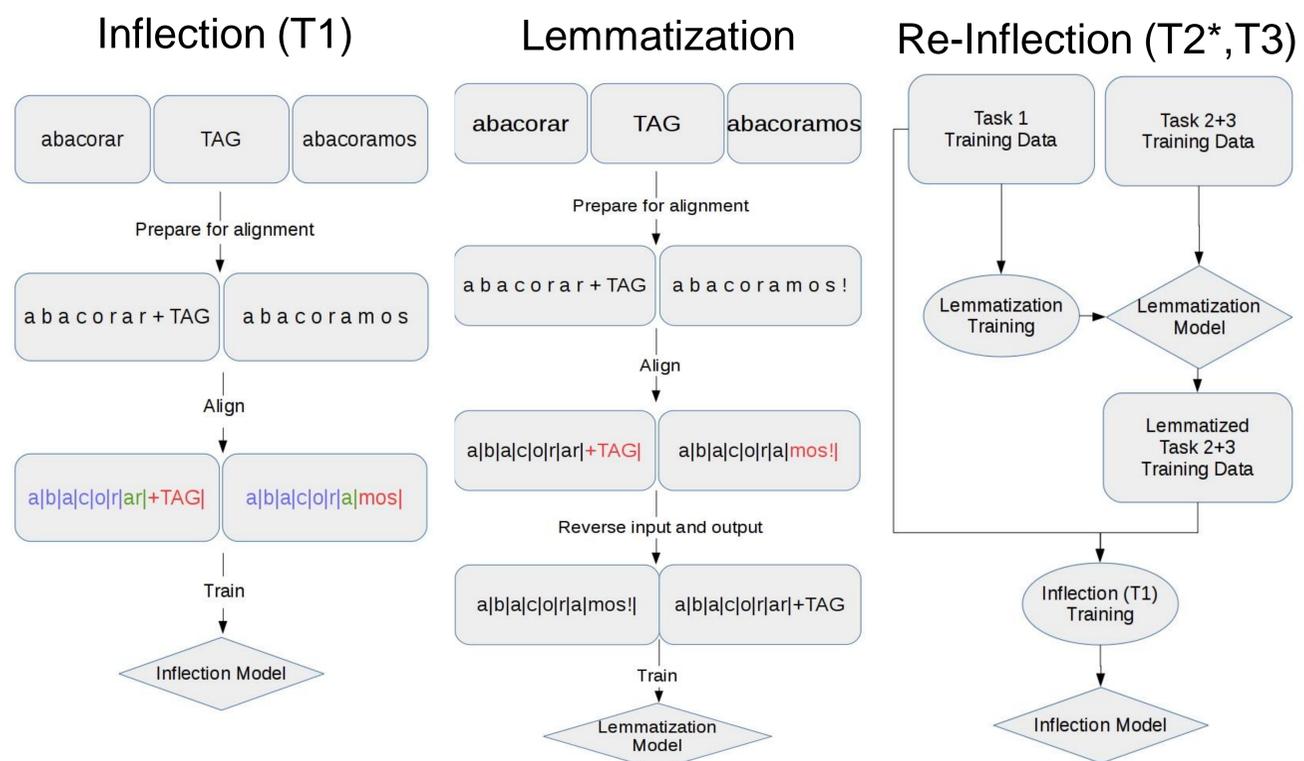
5. Language Specific Methods

Spanish: Diacritical marks indicate irregular stress, (e.g. á in dáras); we treat these marks as distinct symbols (e.g. da'ras)

Finnish, **Turkish**, **Hungarian**: In these agglutinative languages, vowels in stems and suffixes change to share certain phonetic features. We augment DirecTL+ with features to model this.

Georgian: We remove preverb morphemes, which are inconsistent and difficult to predict.

Arabic: Consonants are divided into *sun* and *moon* letters; this distinction affects the definite article "al-"; we correct related errors automatically.

6. Test Results
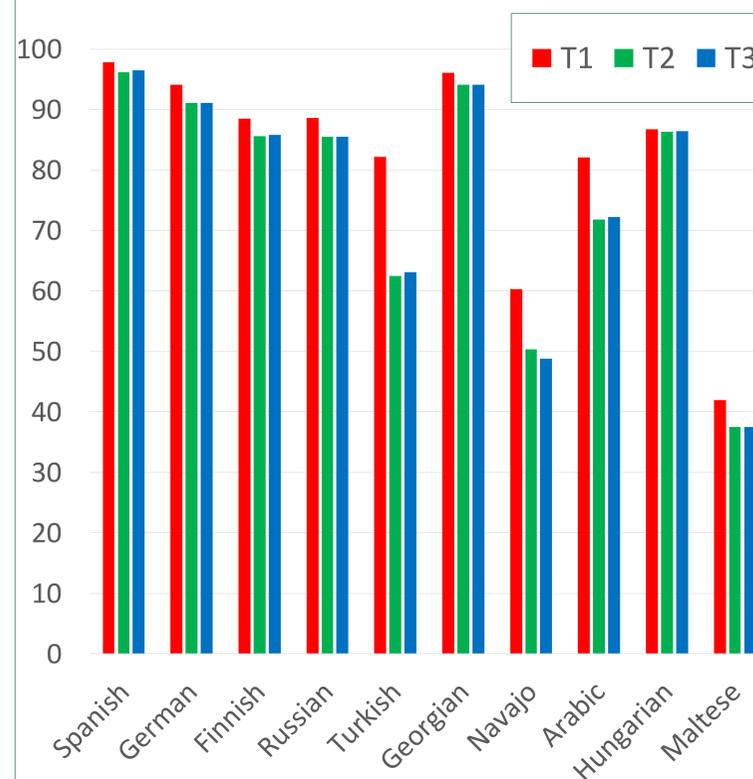
7. Re-ranking with a Text Corpus

Inflect the Russian verb "дождаться", in the future tense, second person, singular, perfective aspect.
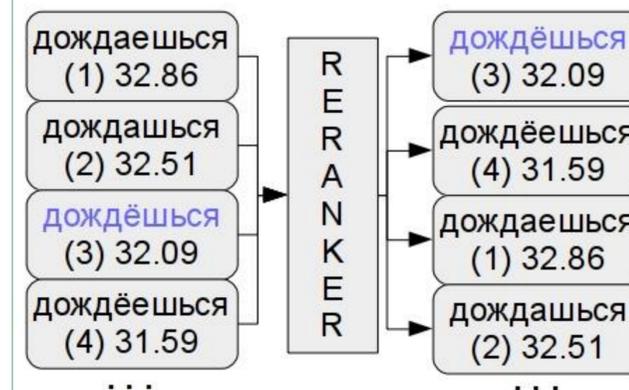
Features:
1. Normalized transduction model score.
2. Presence in a Wikipedia corpus.
3. Normalized language model log-probability.

Usually gives a small improvement.

8. Conclusion

- Previous work has been limited to western European languages.
- String transduction method performs well on various morphologically complex languages and can be adapted to reinflection.
- On the other hand, there is room for improvement on languages like Maltese, which provides motivation for future work.
- Each language presents its own unique challenges; our paper contains language-specific error analysis