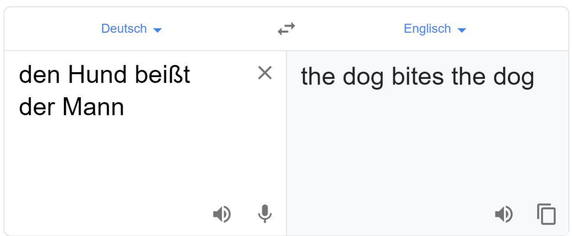
BUILDING A MACHINE TRANSLATOR FOR INFLECTED LANGUAGES

John Coyle

MOTIVATION

- Many of the world‘s languages, particularly Western ones, use morphological makers to specify the function and meanings of words in sentences (German, Russian, Japanese, Arabic, etc)
- Translation with standard SMTs often fails to capture the complexity of these languages:



(correct translation: the man bites the dog)

PROBLEM DEFINITION

- Build a translator that uses knowledge about a language‘s inflection patterns to translate sentences in that language into English (for my project I chose Latin because it has an extremely complex morphological system, because I know it and because all online machine translators do an egregiously poor job of translating)

CHALLENGES

- Machine translation is hard in general
- Extracting morphological features from words (case, tense of verbs, root word
- Cases might have like endings (‘oculis’ in latin can mean ‘with the eyes’ or ‘to the eyes’)

APPROACHES

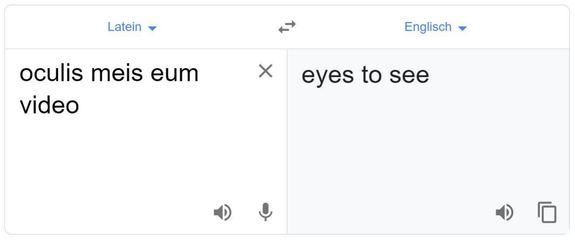
- Breaking individual words into sets of features
- Ex:

„me“ (English) → {*lemma:* ‚I‘, *pos:* ‚noun‘, *case:* ‚acc‘, etc.}

- Translate based on these features individually instead of just the word ‚me‘ e.g.
- Look at the probability of a feature in the latin sentence mapping to a specific english word and include the english word in the translation if it is probable enough
- Reorder english sentence to make more sense

RESULTS

- Example sentence: „oculis meis eum video“ which means „I see him with my eyes“:




```
oculis meis eum video ---------> i see him with eyes my
```

- With more advanced feature extraction can translate correctly words that have never been used in that way e.g. in training data, the word for dog, ‚canis‘, is only used in nominative plural, but it can be translated properly in other cases, and ‚nasus‘, which means nose appears once also in the dative, and ‚vult‘ only appears in first person singular; and yet:

```
canis meus te interficit ---------> my the dog kills you
```

ANALYSIS

- Difficulties with words like „the“ or „a“ or lack of an article:

```
canis meus te interficit ---------> my the dog kills you
```

- Problems with word order sometimes:

```
te amo corde meo ---------> i love you with my heart
```

but,

```
eum video oculis meis ---------> i see him with eyes my
```

Translating with just statistical associations can mean that if one word is translated multiple ways, it doesn‘t get translated at all:

```
amor eius ---------> love
```

For example, „eius“ (see above) can be translated as either ‚his‘, ‚her‘s‘ or ‚its‘, meaning that the word is simply ommitted

NEXT STEPS

- More robust feature parser and more training for figuring out what features to assign to each word
- Better handling of the possibility of multiple possible translations
- Some algorithm for checking the likelihood of sentences which would allow for reordering