Capturing the POS and Entity types of unaligned words

Introduction

I have captured all the unaligned words' POS and Entity types and formulated a summary of what is found right now.

Data Description

I have used EuroParl data now: The data is clean and is updated from Do not Rely on Relay Translations: Multilingual Parallel Direct Europarl.

The dataset has parallel sentences that are collected and separated based on direct sentences or translated ones and also accommodate its native speaker or not. I have used the data from direct native speakers to other language-translated ones.

Currently, I’m using only the data in English and German.

Word Alignment Tool

I have used two main alignment tools. The fast align and eflomal. Since there were a few alignment issues with regards to the fast-align, I have chosen eflomal and also extracted the data from it for getting a summary of the unaligned data.

POS and Entity mapping

The POS tags and entity types were found using the large model from Spacy.

Results and Observations

There were a lot of PROPN which are deemed as unaligned words. This might be because of the cultural references. Because The word “Madam” is used along with the subject when referring to a particular positioned female while translating to English. “Madam” is used as a cultural reference here to honor the position the female member holds.

The Pronouns that have been unaligned are the ones that are necessary for the construction of the language structure and also for preserving the reference to a specific person.

The table below shows that the other POS tags are less unaligned compared to the above ones.

Table 1: Fast-align POS tag summary

5743	PROPN
5218	PRON
2300	ADV
2751	VERB
2638	AUX
3482	DET
1958	ADJ
4144	NOUN

Table 2: EFLOMAL POS tag

PRON	2329
ADV	1606
PROPN	1676
VERB	1513
AUX	1617
NOUN	2763

Regarding Entities, there seem to be no particular entities in the current data. Using some other data would be good in this regard.