Automatic Punctuation Annotation for UD-Annotatrix

Daniel Swanson

Motivation

UD-Annotatrix is a program for annotating syntatic relations within sentences according to the standards of the Universal Dependencies project. Sentences annotated in this way can be used to train machine translation systems and study patterns in the syntaxes of various languages.

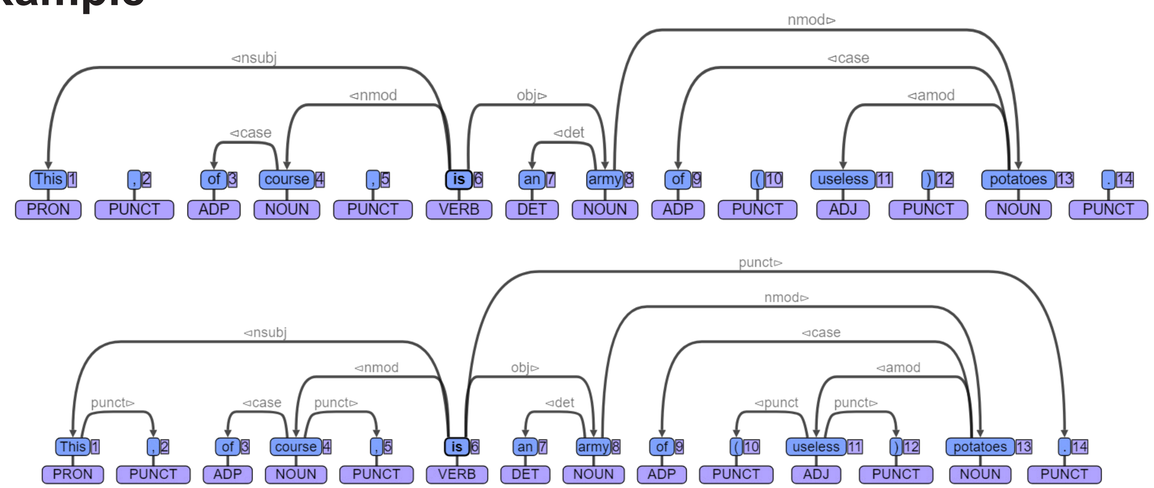
I added functionality to the annotator which automatically handles punctuation marks, for which Universal Dependencies has rules which are consistent across different languages, which should make the annotation process quicker, easier, and potentially more consistent.

Algorithm

In an annotated sentence, every word or punctuation mark is a node in a tree. Each node is either the root of the tree or depends on some other node which is called the "head" (represented by an arrow pointing from the head to the dependent). So the algorithm does the following:

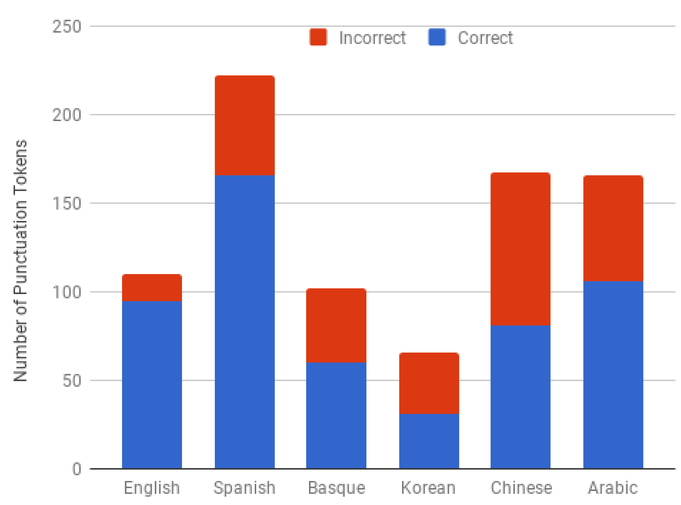
1. Find all the punctuation nodes in the tree
2. For any pairs of quotes or brackets:
    a. They must both attach to nodes that are between them
    b. Try to attach them both to a node that connects to something not between them (the other nodes between are dependent on it)
    c. Otherwise, try the left-most node they both can reach
    d. Otherwise, attach each one to its nearest neighbor
3. For each other punctuation node:
    a. Find every node that it can connect to without crossing any other lines
    b. If it's the last node in the sentence and the root is reachable, attach it there
    c. Otherwise, if it could go with a conjunction, put it there
    d. Otherise, try to attach to the head node of a clause
    e. Otherwise, attach it to the nearest available node

Example



Evaluation

Evaluation was done on a corpus of 300 sentences, 50 from each language. Each sentence had all punctuation connections removed and then reattached by the program. The results were then compared to the original text.



	English	Spanish	Basque	Korean	Chinese	Arabic	Total
Total Connections	110	222	102	66	167	166	833
Correct Connections	95	166	60	31	81	106	539
Accuracy	86.36%	74.77%	58.82%	46.97%	48.50%	63.86%	64.71%

Try it here: http://mr-martian.github.io/ud-annotatrix/index.html

Tyers, Francis M., Maria Sheyanova, and Jonathan Washington. "UD Annotatrix: An annotation tool for universal dependencies." Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories. 2017.