CS 421 – Natural Language Processing – Spring 2018
Term Project (Part 2)

1 General Information

This is the second part of the project. We will deal with syntactic well-formedness in general, and with rudimentary notions of coherence. Due dates are as follows (meant as 11:59pm):

	Due	Points
Competition	4/30 (Mon)	20 (extra-credit)
Part 2	5/7 (Mon)	130

2 Syntactic well-formedness

You will use one of the three parsers (Stanford, OpenNLP, NLTK) from Part 1 to complete the evaluation of the grammatical well-formedness of the essay. In particular, we have not dealt with evaluation criterion *c.iii. sentence formation* (please see p. 2 in the handout for the first part of the project).

We want our students of English to write complete sentences. A criterion for complete sentences could include the following subcriteria:

1. are main sentences formed properly? i.e. they should begin and end properly; the constituents should be formed properly: are there missing words or constituents (prepositions, subject, object etc.)?

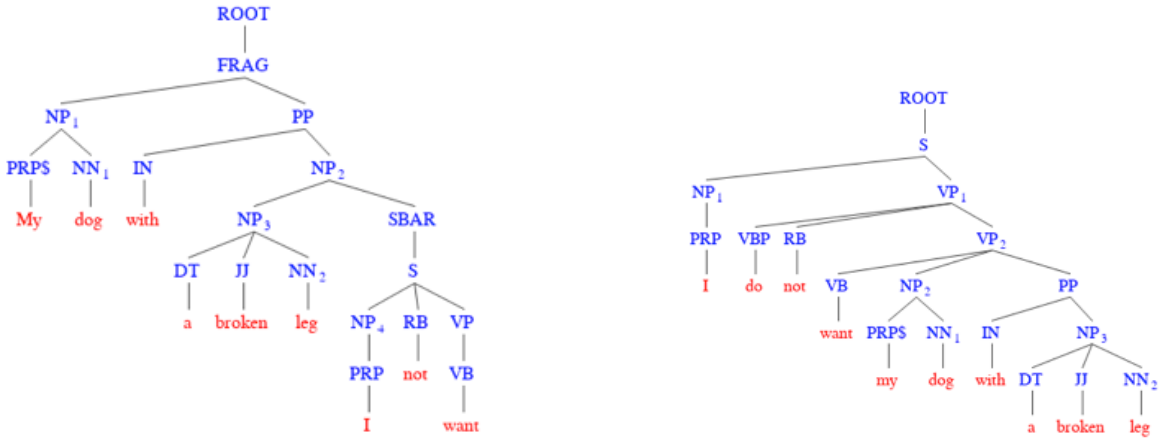
2. If subordinating conjunctions are used (e.g., *when, although, if*), is there a main verb, or a gerund, to go with them? For many subordinating conjunctions, the corresponding clause can be finite, i.e., it has a main verb: *when I travel alone, ...* ; or it's not finite, and includes a gerund: *when traveling alone, ...* But e.g. *because* must be used with a finite verb. You can say *Because I travel alone, ...* , but not *Because traveling alone, ...* Note that we are talking about *because* as a *subordinating* conjunction, not when it's used within an NP like in *Because of my extensive travels, I have lots of frequent flyer miles*. This is one single main clause, there is no subordinate clause.

You can use POS tagging to evaluate some features contributing to *c.iii*, for example correct word order, e.g. in English no main verb should be in first position in a declarative sentence; or no subordinating conjunction should be in a sentence with only one main verb, like in *Because I think the sience and tecnology are developping.* (this is the whole sentence). In general, you are asked to use the results the parser returns to judge criterion *c.iii*. Note that *c.iii* applies to the whole essay, like the other criteria; a single wrong sentence should not drive the score down a lot, if the other sentences are correct.

2.1 How to use parse trees

All these parsers will return at least some parse trees for most (?all) sentences; you can use patterns from the trees to recognize mistakes. For example, consider the incorrect sentence S_1*:"My dog with a broken leg I not want"*, and its correct counterpart S_2*:"I do not want my dog with a broken leg"*.

Figure 1 shows the two parse trees side by side, S_1 on the left, S_2 on the right. These are parse trees returned by the Stanford parser (on its own or as embedded within NLTK), parse trees returned by OpenNLP are very similar.



(a) S_1 (incorrect)

(b) S_2 (correct)

Figure 1: Two parse trees

You will note that the parse tree for S_1 includes a "FRAG" node (for "FRAGment"): this is an indication that the sentence is not complete; additionally, there is no main clause (no *S*) but only a subordinate clause, indicated by the SBAR node. SBAR nodes can be daughters of NPs when they represent a relative clause (see below), but then there must be a main verb denoting the main clause. The tree for the correct S_2 includes an S, but neither FRAG nor SBAR.

One approach may be to write some patterns that correspond to common mistakes, as noted above (e.g. Stanford parser: there's FRAG, and an SBAR without an S), and check whether the output of the parser includes those patterns. In theory, another approach would be to retrain these parsers with the new data, but the corpus of essays is too small; besides, they would have to be annotated with syntactic trees, at least for the Stanford parser.

These are some *possible* ideas – meaning, once you think about it, you may come up with other ways of using the parsers. **Be creative!** In addition, note that the work you did in part 1 to evaluate criteria *c.i* and *c.ii* can also be used to seed the work done by the parser; and conversely, the results of parsing may be useful to better evaluate *c.i* and *c.ii*. You are not asked to redo the work you did in part 1 for these three criteria, but to integrate the two approaches together if they provide complementary information.

2.1.1 More on SBARs – subordinate clauses

Since above we pointed out that SBAR is a signal for the incorrect S_1, we hope you don't misunderstand this as a signal of error per se. Here is some further information about the contexts in which an SBAR, which

represents a subordinate clause, may appear.

An SBAR can be introduced by a subordinating conjunction such as *when, because, although* (marked by IN as POS tag). For example, Figure 2 shows the parse tree for the correct S_3 *"I came because he was sick"*. Note that the SBAR node is a daughter of a VP, and has an immediate daughter which is the conjunction in question.

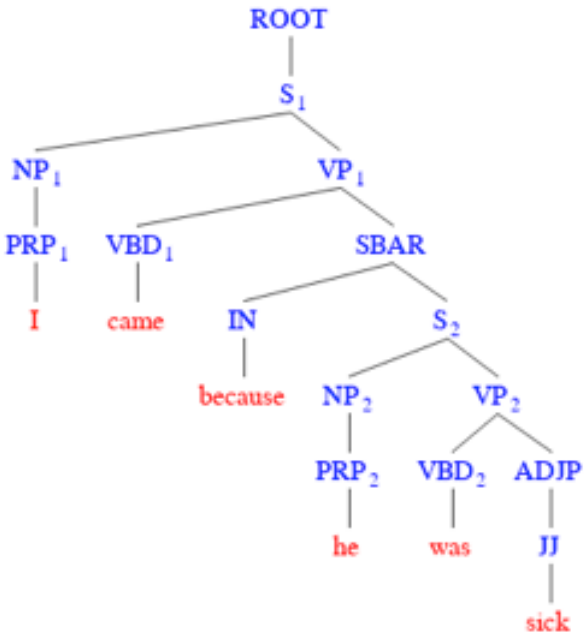


Figure 2: Parse tree for S_3

An SBAR node can also be a daughter of a VP with verbs of *saying* such as *say, tell, claim, report, etc....* These verbs take a whole other sentence as argument, introduced by *that*, as in S_4*: "He claimed that aliens landed in his garden"* (parse tree left as an exercise).

Finally, an SBAR node can be used to denote a relative clause, as in S_5*: The dog I saw was a German shepherd*. Figure 3 shows the parse tree for S_5.

3 Text coherence

The grading criteria we listed in part 1 include *d.i* and *d.ii*, which pertain to semantics and pragmatics, and are repeated here:

d.i Is the essay coherent?

d.ii **(For graduate students only)** Does the essay answer the question / address the topic?

3.1 Essay Coherence via pronouns

d.i corresponds to an open research problem in NLP. However, it is possible to come up with a coarse

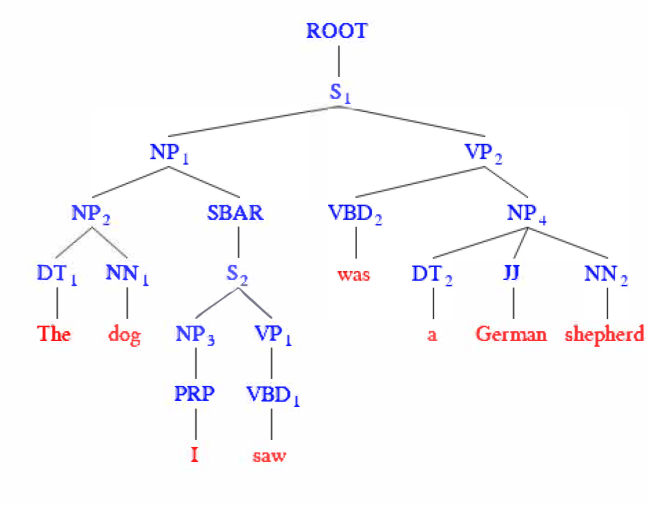


Figure 3: Parse tree for S_5

measure of coherence via referential expressions, which is still too complex in general for these essays: many referring expressions are definite NPs (like *the places you choose to go to*) which are too difficult to deal with within a project like this. However, we can get some mileage out of assessing whether pronouns are used felicitously.

Here are some observations on pronoun usage in these essays.

1. First person singular pronouns and possessive adjectives *I, me, my, mine* refer to the speaker / writer, are solved based on who the speaker is, and are not ambiguous. Same for first person plural pronouns, *we, our*, although they are harder to interpret since they refer to a group that includes the speaker. Second person pronouns and possessive adjectives (*you, your*) can be used as well, in an impersonal sense, as in the following example from the excerpt of the *high* essay included in the first part of project: ... *going to the places you choose to go to and discovering everything on your own*.

2. Third person singular pronouns are hardly used in these essays. Doublecheck if they do. If you find a *he* or *she* you can quickly assess whether it is used properly: any third person pronoun should have a possible antecedent. If *she* is used and no feminine entity has been introduced, then *she* is wrong (see below a note on where to find the information about gender and number); likewise for *he* and male antecedents.

3. On the other hand, third person plural pronouns (*they*) are often used. For these,

   (a) First, you should check if there are potential correct antecedents: either plural nouns, or nouns with compatible number (see sec 21.6.4 in the book), but used properly. Ie, *someone, group, family* can be used as antecedents for *they/them*, but it often doesn't sound felicitous when the antecedent is in a prepositional phrase:

   - *A group travelling together can be fun. You will get to know them*: the pronoun *them* is felicitous

- *I don't agree that the best way to travel is in a group. They will have many problems*: the pronoun *they* is not as felicitous

(b) Second, the antecedent to *they/them* should not be too far: so, a pronoun should have an appropriate referent in the previous one-two sentences; the referent could be another pronoun referring to the same entity. This is called a chain.

(c) Finally, if there is more than one possible antecedent, one of them should be more prominent than the other. In our simplified scenario, more prominent means it has been mentioned more recently; the further apart the various possible antecedents are, the better the referent is.

3.1.1 One algorithm, more in detail

One possible algorithm would be along the following lines:

1. Collect all pronouns and possessive adjectives
2. Eliminate all pronouns/adjectives that are not third person
3. For singular third person pronouns and possessives *he/she/his/her*, check the existence of appropriate male/female antecedents as mentioned earlier.
4. For plural third person pronouns and possessives *they/them/their*, check if there are possible antecedents:
   (a) no plural antecedent, or no singular antecedent with compatible number: mistake. If antecedent is singular, try to assess if it's used correctly (given its syntactic position)
   (b) only one possible antecedent: correct / felicitous
   (c) more than one possible antecedent: evaluate how ambiguous that pronoun is, based on how many antecedents there are, and how recent they are. The more ambiguous, the less felicitous.

For the last step, you can use the results that the coreference modules in the Stanford / OpenNLP / NLTK packages provide. However, what these packages return is one possible interpretation for pronouns (and possibly other NPs), i.e. they don't tell you how many possible antecedents there are for a single pronoun. If you use the results of these coreference modules, you will have to augment them, as just discussed.

Notes

1. You can exploit POS tags for number.
2. Note that *they* sometimes does not have an explicit previous plural referent, but rather, refers to collections of items, as in *I have a son and a daughter. They play together a lot.* You should think of some appropriate heuristics to deal with these cases (if they arise).
3. As usual, we are greatly oversimplifying the problem. For example, it is not true that if there's no antecedent within the previous 1-2 sentences the pronoun is infelicitous, it really depends on the global structure of discourse.

4 Topic Coherence (for graduate students)

For criterion *d.ii* (Does the essay address the topic?), we want to understand whether the student has written about the stated topic. For example, suppose our topic was *will people have fewer kids in the future?*: you could see how many of the common nouns in the essay are related to family, but also to potential relevant topics, such as work, economics, etc.

To assess this, you could check the common nouns in the essay by exploiting Wikipedia, or electronic dictionaries, eg Wordnet (http://wordnet.princeton.edu) or ontologies, e.g. SUMO (http://www.ontologyportal.org/). In Wordnet, words are organized in terms of hypernyms (is-a relation), meronym (part-of) and many other relations. You can exploit that information to find related words to the ones in the essay topic. The problem that you may encounter is that there are many senses for a word and one would need to disambiguate the correct sense. But for your project you can assume that we know the sense and just use the correct sense.
In SUMO, one subhierarchy is on "familyRelation". For example, if you browse *aunt* in SUMO, you get

```
109823502 the sister of your father or mother; the wife of your uncle.

   SUMO Mappings: familyRelation (subsuming mapping)
```

Once you have an estimate for how many proper / common nouns refer to family, geography, work, and other potential relevant topics, you can assess the relevance of the essay in its entirety – for example, by computing the percentage of all the nouns that relate to the relevant subtopics with respect to all the nouns in the essay.

This is a fairly naive method. To start with, it doesn't take verbs into account. If you have more creative ideas, feel free to experiment. Creativity, not only results, will be rewarded when grading your work (as long as it makes sense).

5 Report and Submission

When submitting your work, please follow the guidelines for the first part of the project. Specifically, your submission will now contain the complete project and updated files.

You are also required to write a project report. In the report you have to comment on your project, what worked, what did not work, what you learned from the assignment, what you think would be necessary to take your software to a higher level of performance, etc. The report should be at least one page long – ie, not just a single paragraph.