Clausal Complement? Passive Subject? Grammatical Relationships for (non) Linguists

First Lastname

Firstname Lastname

ABSTRACT

INTRODUCTION

What does X do? How is Y described? Most analysts and researchers ask themselves these questions at one point or another while trying to understand a subject. So far, the field of information retrieval has tackled this problem indirectly. The analyst enters some kind of query, and the system returns some results. When the underlying data has structure, queries can be specific and targeted – and there is now a substantial body of work [cite] on how best to issue structured queries over *structured* data. But what if the "data" is is text – medical literature, legal records, interview transcripts? Although text is richly structured by rules of grammar and narrative, it is rarely treated as such. While there is a lot of work on extracting various kinds of structured information from text, there is little on how to make it easy for users to access and query over it. As the questions above show, however, structured queries over language can be extremely useful. In grammatical terms, those two questions become "What are the verbs of which X is the subject?" "What are the adjectives that modify Y?"

Adapting existing structured-query interfaces to grammatical search is problematic for several reasons. The first is that the structures in language are not explicit, like columns in a table, but are implicit, and have to be extracted computationally. Only in the last decade have computational linguistic technologies become fast and accurate enough to use in the real world.

The second problem is the lack of programming experience among searchers. Current structured query like SPARQL have complex syntaxes that require time and effort to learn. For example, here is a SPARQL query for "What are all the country capitals in Africa?":

```
SELECT ?capital ?country
WHERE {
  ?x cityname ?capital ;
     isCapitalOf ?y .
  ?y countryname ?country ;
     isInContinent Africa .
}
```

Even if we assume that only professional analysts engaging in information-intensive work would want to issue such targeted queries, programming experience is scarce. One study on a query language for grammatical structures from sentences found that that only 50% of the people who wanted to use it had programming experience [].

The third problem is that there are no common-language terms for grammatical relationships even though ordinary people are perfectly capable of understanding and using them. Modern parsers use standard linguistics terminology to label their outputs, but those technical names and definitions are not always accessible to those outside the field. Take the phrases "he threw the ball" and "the ball was thrown by him". In both cases, it is clear that 'he' is is the one who 'threw' whereas, in grammatical terms, the first phrase is in the active case, and the phrase is in the passive case. The Stanford Dependency Parser [], for example, outputs two different variants of the verb-subject relation: `nsubj(he, threw)` and `nsubjpass(he, threw)`. A grammatical search system therefore has to bridge the gap between the relations that are recognizable to people and the relations that are extracted from the data.

We conducted experiments to investigate how grammatical relationships between words can be made more recognizable. Following the principle of recognition over recall, we hypothesized that examples would help people identify grammatical relationships more accurately rather than formal terms.

We tested two types of examples: a list of matching words, and a list of matching phrases containing the relationship. This is because a grammatical relationship has three components: the two words that enter into the relationship, and the relationship itself. The words are explicitly visible in the text, and the grammatical relationship is derived from contextual information such as the part of speech of the verb, the relative ordering, and any accompanying words.

We chose Amazon's Mechanical Turk crowdsourcing platform as a source of study participants. The ability to query over grammatical relationships has applications in many fields outside linguistics and language study, so we wanted to avoid having study participants with specialized backgrounds.

We presented participants with a series of identification tasks. In each task, there was a 'query' word and list of sentences containing a grammatical relationship between that query word and some other words. Their task was to pick the correct relationship from list of four choices. We presented the choices in three different ways – as a short label using linguistic terminology, a linguistic label accompanied by a list of example words that matched, and a linguistic label accompanied by a list of phrases in which that relationship surfaced.

Our results confirm that showing examples significantly improves the accuracy with which grammatical relationships are recognized.

Our hypothesis was the following:

> H1. Grammatical relations can be made more recognizable by showing examples of words or phrases that match.