Цель: сравнить работу парсеров, используя учебный корпус для подбора примеров предложений, обладающими различными грамматическими особенностями, а также найти самый удобный и быстрый парсер, отвечающий всем или большинству требований.

Описание работы:

1. Пояснение устройства парсера: сложность алгоритмов и т.д
2. Пояснение логики работы парсера с примерами, которые могли бы показать, как парсер работает с:
    - Подлежащим, если у него есть/нет зависимых слов
    - Сказуемым (точно так же)
    - Однородными члены
    - Разбиением на клаузы
    - Сложными случаями
3. Повторение пункта 2, но с опечатками в предложениях двух видов: ***±*** ***пунктуация и её отсутствие***
    - случай,когда при замене/пропуска одной буквы у нас существующее слово меняется на существующее (they → the/thy, и что-то похожее)
    - случай,когда существующее слово становится несуществующим (they -> th)
4. Повторение пункта 2, но проверкой влияния наличия/отсутствия знаков препинания на деление клауз, а также определение синтаксических связей внутри предложений.

Ниже можно увидеть таблицу, в которой наглядно представлено сравнение работы парсеров на конкретных примерах с разными грамматическими особенностями. Таблица поделена на условные разделы, которые нам интересны. Она описывает общую логику работы парсеров (пункт 2).

<table>
<tr><th></th><th>UDPipe</th><th>Stanford CoreNLP</th><th>SpaCy</th></tr>
<tr><td rowspan="2">Подлежащее, если у него есть/нет зависимых слов

a. It leads to one problem.
b. The number of conflicts decreases.
c. Sport facilities are a good way to solve a problem.</td><td>a. It (nsubj)
b. number (nsubj) of conflicts (nmod)
c. Sport (compound) facilities (nsubj)</td><td>a. It (nsubj)
b. number (nsubj) of conflicts (nmod)
c. Sport (compound) facilities (nsubj)</td><td>a. It (nsubj)
b. Justice (for authors - зависит от nsubj)
c. Sport (зависит от nsubj) facilities (nsubj)</td></tr>
<tr><td colspan="3">Подлежащее -- nsubj, зависят от него modifiers и compounds</td></tr>
<tr><td rowspan="2">Сказуемое (точно так же)

a. In 21 century science goes forward.
b. They have caused some huge problems too.</td><td>a. demonstrates (root)
b. have (aux) caused (root)</td><td>a. The figure demonstrates (root)
b. They have (aux) caused (root)</td><td>a. goes. (science - nsubj)
b. have caused (caused - root, they-nsubj)</td></tr>
<tr><td colspan="3">Смысловой глагол -- root, вспомогательные элементы -- aux, везде определены верно.</td></tr>
</table>

<table>
<tr><td>Однородные члены

a. Non-metallic minerals and chemicals have approximately 10 percent of rail transportation.

b. Non-metallic minerals, chemicals have approximately 10 percent of rail transportation.

c. It also develops and strengthens its economy.</td><td>a. minerals (nsubj) and (cc) chemicals (conj)

b. minerals (obl) , (punct) chemicals (nsubj)</td><td>a. minerals (nsubj) and (cc) chemicals (conj)

b. minerals (nsubj) , (punct) chemicals (appos)</td><td>a. minerals (nsubj) and(cc) chemicals(conj) (have - root)
b. Non-metallic minerals, chemicals (разделяет на две разные клаузы по запятой) - нужно проверить на других примерах похожих</td></tr>
<tr><td></td><td>c. develops (root) and (cc) strengthens (conj)</td><td>c. develops (root) and (cc) strengthens (conj)</td><td>c. develops(root) and (cc) strengthens (conj)</td></tr>
<tr><td></td><td colspan="3">У Spacy и Stanford однородный член присоединяется к первому на том же уровне, что и соединительный союз и обозначается как conj (получается два уровня), если есть запятая, Spacy не справляется, так как делит сразу предложение на разные клаузы. UDPipe присоединяет союз на уровень ниже второго однородного слова (получается три уровня). Только Stanford правильно работает с однородными членами, разделенными запятой, без союза.</td></tr>
<tr><td>Разбиение на клаузы

a. relative clauses
The metals which took the first place in the previous chart take only the 5th with 11%.

b. subordinate clauses
A lot of people think that it is ugly and unimpressive.

c. coordinate clauses
Children are our future, and they should be healthy from birth.

d. adverbial clause modifier
Maybe I would have gone to France for several months and would speak French as fluently as the French do.</td><td>a. which (nsubj) took (acl:relcl) the first place in the previous chart</td><td>a. which (nsubj) took (acl:relcl) the first place in the previous chart</td><td>a. which(nsubj) took (relcl, зависит от metals-nsubj, зависит от took-root) the first place in the previous chart .</td></tr>
<tr><td></td><td colspan="3">Relative clauses всеми парсерами распознаются, вершина зависимой клаузы зависит от слова в главной клаузе, к которому семантически присоединено.</td></tr>
<tr><td></td><td>b. that (mark) it (nsubj) is (cop) ugly (ccomp) and (cc) unimpressive (conj)</td><td>b. that (mark) it (nsubj) is (cop) ugly (ccomp) and (cc) unimpressive (conj)</td><td>b. that(mark) it (nsubj) is(ccomp, зависит от root - think) ugly(acomp) and unimpressive (conj - зависит от ugly).</td></tr>
<tr><td></td><td colspan="3">Subordinate clauses всеми парсерами распознаются, вершина зависимой клаузы зависит от вершины главной.</td></tr>
<tr><td></td><td>c. and (cc) they (nsubj) should (aux) be (cop) healthy (conj)</td><td>c. and (cc) they (nsubj) should (aux) be (cop) healthy (conj) from birth</td><td>c. and (cc) they should(aux) be(conj, зависит от root - are) healthy from birth.</td></tr>
<tr><td></td><td colspan="3">Coordinate clauses всеми парсерами распознаются, вершина зависимой клаузы зависит от вершины главной.</td></tr>
<tr><td></td><td>d. as (case) the (det) French (obl) do(advmod)

Не распознает такое как отдельную клаузу: считает все</td><td>d. as (mark) the French (nsubj) do (advcl)</td><td>d. as the French(nsubj) do(advcl - зависит от fluently)</td></tr>
</table>

	сравнительным оборотом без подлежащего и сказуемого		
	С последним пунктом проблема возникла у UDPipe, остальные парсеры распознали конструкцию и верно выделили подлежащее и сказуемое в интересующей клаузе.		
5. сложные случаи **a. инверсия** The bachelor's degree have a 45 per cent of males and 55 per cent of females.	a. degree (nsubj), have (root)	a. degree (nsubj), have (root)	a. degree (nsubj), have (root)
	Все парсеры не отслеживают нарушение порядка слов SVO вне зависимости от согласования, следовательно, не улавливают инверсию, даже если определенные подлежащее со сказуемым не согласовываются		
b. gerund as a subject Having thoughts of committing crimes comes from the period of socialization.	b. Having *(csubj)* thoughts of committing crimes comes *(root)*	b. Having *(csubj)* thoughts of committing crimes comes *(root)*	b. Having *(csubj)* thoughts of committing crimes comes *(root)*
	Верно определены подлежащее-герундий и сказуемое всеми парсерами.		
c. there is/are There are several opinions on this issue that I am going to discuss.	c. There *(expl)* are *(root)* several opinions ***(nsubj)***	c. There *(expl)* are *(root)* several opinions ***(nsubj)***	c. There(*expl*) are(*root*) opinions**(*attr*)** *Именная часть подлежащего выделяется тегом attr*
	Верно определяются подлежащее и сказуемое клауз в конструкциях такого типа всеми парсерами.		
d. предложения с именным сказуемым Sport facilities are a good way to solve a problem.	d. Sport facilities *(nsubj)* are ***(cop)*** a good way ***(root)***	d. Sport facilities(*nsubj*) are ***(cop)*** a good way ***(root)***	d. Sport facilities(*nsubj*) are **(*root*)** a good way **(*attr*)**
	Из-за разной логики парсеров по-разному выделяются вершины клауз и, следовательно, зависимости внутри них. UDPipe и Stanford считают сказуемым именную часть, а глагол-связку обозначают как copula. Spacy же считает сказуемым глагол-связку, а именную часть обозначает как attribute.		
e. пассив The building was destroyed by the demolition crew.	e. The building *(nsubjpass)* was *(auxpass)* destroyed *(root)*	e. The building *(nsubjpass)* was *(auxpass)* destroyed *(root)*	e. The building(*nsubjpass*) was(*auxpass*) destroyed(*root*)
	Конструкция пассива верно определяется всеми парсерами, предусмотрены отдельные теги для подлежащего и сказуемого в пассивных конструкциях.		
f. have-иметь и have-вспомогательный Our society always has crime. I have done my homework.	f. society *(nsubj)* has *(aux, AUX)* crime *(root, VERB)* I *(nsubj)* have *(aux, AUX)* done *(root)* *В идеале, должен различать have-aux и*	f. society *(nsubj)* has *(root, VERB)* I *(nsubj)* have *(aux, AUX)* done *(root)*	f. has(*root*, но в частеречном разборе считает have aux в любом случае) society (*nsubj*)

	have-verb, однако в значительной части случаев(см f.1) не справляется с этой задачей		
	Для Spacy have в частеречном разборе всегда AUX, но это не влияет на проверку правильного/неправильного согласования, так как, когда Spacy выделяет have как root - have - модальный или вспомогательный, а иначе have - вспомогательный, потому что логика парсера такова, что в аналитических формах root является смысловой глагол. *Для остальный парсеров ситуация иная: в UDPipe есть разделения на have-aux и have-verb, однако в значительной части случаев(см f.1) оно не является верным. У Standford NLP также есть разделения на have-aux и have-verb. Но стоит отметить, что это разделение нужно для определения вершины клаузы, а это все парсеры делают верно.*		

English version:

	UDPipe	Stanford CoreNLP	SpaCy
Subject if it has / doesn't have dependent words a. It leads to one problem. b. The number of conflicts decreases. c. Sport facilities are a good way to solve a problem.	a. It *(nsubj)* b. number *(nsubj)* of conflicts *(nmod)* c. Sport *(compound)* facilities *(nsubj)*	d. It *(nsubj)* e. number *(nsubj)* of conflicts *(nmod)* f. Sport *(compound)* facilities *(nsubj)*	d. It (*nsubj*) e. Justice (for authors - зависит от *nsubj*) f. Sport (зависит от *nsubj*) facilities (*nsubj*)
	Subject is marked as 'nsubj', dependent words are modifiers or compounds.		
Predicate **a. In 21 century science goes forward.** b. They have caused some huge problems too.	a. demonstrates *(root)* b. have *(aux)* caused *(root)*	a. The figure demonstrates *(root)* b. They have *(aux)* caused *(root)*	**a. goes. (*science - nsubj*)** b. have caused (*caused - root, they-nsubj*)
	Main verb is marked as 'root'; auxiliary verbs are marked as 'aux'. All 3 parsers marked them correctly		
Homogeneous parts a. Non-metallic minerals and chemicals have approximately 10 percent of rail transportation.	a. minerals *(nsubj)* and *(cc)* chemicals *(conj)* b. minerals ***(obl)*** , *(punct)* chemicals ***(nsubj)***	a. minerals *(nsubj)* and *(cc)* chemicals *(conj)* b. minerals ***(nsubj)*** , *(punct)* chemicals ***(appos)***	a. minerals (*nsubj*) and(*cc*) chemicals(*conj*) (*have* - root) **b.** Non-metallic minerals, chemicals

<table>
<tr><td rowspan="2">b. Non-metallic minerals, chemicals have approximately 10 percent of rail transportation.

c. It also develops and strengthens its economy.</td><td>c. develops (root) and (cc) strengthens (conj)</td><td>c. develops (root) and (cc) strengthens (conj)</td><td>c. develops(root) and (cc) strengthens (conj)</td></tr>
<tr><td colspan="3">Spacy and Stanford CoreNLP place the head of the second homogeneous part on the same level as the coordinating conjunction. Both words are connected to the head of the first homogeneous part, which is placed on one level above. Thus two levels are made.

UDPipe connects the coordinating conjunction to the second homogeneous part (which itself is connected to the first homogeneous part), thus making three levels.

Spacy doesn't parse well sentences in which homogeneous parts are divided by a comma because it often separates clauses and creates two trees, one for each clause.

Stanford CoreNLP handles well homogeneous parts that are divided by a comma.

UDPipe does not recognise two homogenous parts divided by commas. The parser marks one of the parts correctly, while the second one depends from it either as oblique (for nouns) or as adverbial clause (for verbs). In case with three or more comma-separated parts UDPipe performs well.</td></tr>
<tr><td rowspan="8">Dependent clauses

a. relative clauses
The metals which took the first place in the previous chart take only the 5th with 11%.

b. subordinate clauses
A lot of people think that it is ugly and unimpressive.

c. coordinate clauses
Children are our future, and they should be healthy from birth.

d. adverbial clause modifier
Maybe I would have gone to France for several months and would speak French as fluently as the French do.</td><td>a. which (nsubj) took (acl:relcl) the first place in the previous chart</td><td>a. which (nsubj) took (acl:relcl) the first place in the previous chart</td><td>a. which(nsubj) took (relcl, зависит от metals-nsubj, зависит от took-root) the first place in the previous chart .</td></tr>
<tr><td colspan="3">All parsers correctly recognize relative clauses. The head of the relative clause is dependent from the word of the main clause to which it is semantically related to.</td></tr>
<tr><td>b. that (mark) it (nsubj) is (cop) ugly (ccomp) and (cc) unimpressive (conj)</td><td>b. that (mark) it (nsubj) is (cop) ugly (ccomp) and (cc) unimpressive (conj)</td><td>b. that(mark) it (nsubj) is(ccomp, зависит от root - think) ugly(acomp) and unimpressive (conj - зависит от ugly).</td></tr>
<tr><td colspan="3">Subordinate clauses are recognized correctly by all parsers. The head of the subordinate clause is dependent from the head of the main clause.</td></tr>
<tr><td>c. and (cc) they (nsubj) should (aux) be (cop) healthy (conj)</td><td>c. and (cc) they (nsubj) should (aux) be (cop) healthy (conj) from birth</td><td>c. and (cc) they should(aux) be(conj, зависит от root - are) healthy from birth.</td></tr>
<tr><td colspan="3">Coordinate clauses are recognized correctly by all parsers. The head of the coordinate clause is dependent from the head of the main clause.</td></tr>
<tr><td>d. as (case) the (det) French (obl) do(advmod)

UDPipe failed to recognize this clause, marking it as an adverbial modifier instead.</td><td>d. as (mark) the French (nsubj) do (advcl)</td><td>d. as the French(nsubj) do(advcl - dependent from 'fluently')</td></tr>
</table>

<table>
<tr><td></td><td colspan="3">The only parser that failed to recognize comparative clause was UDPipe. Spacy and StanfordNLP performed the task correctly.</td></tr>
<tr><td rowspan="10">5. difficult cases

a. inversion
The bachelor's degree have a 45 per cent of males and 55 per cent of females.

b. gerund as a subject
Having thoughts of committing crimes comes from the period of socialization.

c. there is/are
There are several opinions on this issue that I am going to discuss.

d. sentences with a noun predicate
Sport facilities are a good way to solve a problem.

e. passive
The building was destroyed by the demolition crew.

f. semantic ‘have’ and auxiliary ‘have’
Our society always has crime.

I have done my homework.</td><td>a. degree (nsubj), have (root)</td><td>a. degree (nsubj), have (root)</td><td>a. degree (nsubj), have (root)</td></tr>
<tr><td colspan="3">All three parsers do not recognize the disruption of the standard SVO order regardless of whether the agreement is right or wrong. Therefore, they are unable to recognize inversion even if certain subjects and predicate verbs do not have correct agreement.</td></tr>
<tr><td>b. Having (csubj) thoughts of committing crimes comes (root)</td><td>b. Having (csubj) thoughts of committing crimes comes (root)</td><td>b. Having (csubj) thoughts of committing crimes comes (root)</td></tr>
<tr><td colspan="3">Gerund subject is marked as ‘csubj’. All three parsers recognize it correctly.</td></tr>
<tr><td>c. There (expl) are (root) several opinions (nsubj)</td><td>c. There (expl) are (root) several opinions (nsubj)</td><td>c. There(expl) are(root) opinions(attr)
Именная часть подлежащего выделяется тегом attr</td></tr>
<tr><td colspan="3">All three parsers are able to correctly parse this type of sentence.

All three parsers mark ‘there’ as ‘expl’, ‘are’ or ‘is’ as ‘root’. UDPipe and Stanford CoreNLP mark the noun as ‘nsubj’; Spacy marks the noun as ‘attr’.</td></tr>
<tr><td>d. Sport facilities (nsubj) are (cop) a good way (root)</td><td>d. Sport facilities(nsubj) are (cop) a good way (root)</td><td>d. Sport facilities(nsubj) are (root) a good way (attr)</td></tr>
<tr><td colspan="3">Spacy differently marks the head of the main clause because its parsing logic is different from that of UDPipe and Stanford CoreNLP. The latter two mark the noun predicate as ‘root’, and the copula verb as ‘cop’. Spacy, on the other hand, marks the noun predicate as ‘attr’, and the copula verb as ‘root’.</td></tr>
<tr><td>e. The building (nsubjpass) was (auxpass) destroyed (root)</td><td>e. The building (nsubjpass) was (auxpass) destroyed (root)</td><td>e. The building(nsubjpass) was(auxpass) destroyed(root)</td></tr>
<tr><td colspan="3">Passive clauses are recognized correctly by all three parsers. There are distinct tags for auxiliary verbs and subjects in the passive construction – ‘auxpass’, ‘nsubjpass’.</td></tr>
<tr><td>f. society (nsubj) has (aux, AUX) crime (root, VERB)

I (nsubj) have (aux, AUX) done (root)

Ideally, UDPipe should distinguish between semantic ‘have’ and auxiliary ‘have’, but it often (see f.1) fails to do so.</td><td>f. society (nsubj) has (root, VERB)

I (nsubj) have (aux, AUX) done (root)</td><td>f. has(root, AUX) society (nsubj)</td></tr>
</table>

	Both UDPipe and Stanford CoreNLP have a distinction between semantic 'have' and auxiliary 'have': their POS-taggers mark the former as 'VERB' (while also syntactically marking it as 'root') and the latter as 'AUX' (while also syntactically marking it as 'aux'); . However, in most cases UDPipe marks the verb (see f.1) incorrectly. Stanford CoreNLP usually handles the distinction well. *Spacy's POS-tagger always marks 'have' as 'AUX', but it does not disrupt the work of the parser. In most cases, when the verb is semantic it is marked as 'root', and when the verb is auxiliary it is marked as 'aux'.*		

It also develops, strengthens its economy.

Asia, Europe, Africa didn't change during two periods.