NLP-Werkstatt
SS2017
Extraktion verschiedener Annotationsebenen
Hausaufgabe – Abgabe bis 10.05.2017, 14:00 Uhr

Gegeben sei eine Datei aus den *OntoNotes*-Daten[1] (*cctv_0001.v4_gold_conll*). Die Datei ist mit der Aufgabenbeschreibung herunterzuladen unter http://www.cl.uni-heidelberg.de/courses/ss17/annotierteKorpora/material/cctv_0001.v4_gold_conll.

Der Text ist mit verschiedenen Annotationsebenen annotiert. Eine Beschreibung des CoNLL-Formats findet man hier: http://conll.cemantix.org/2012/data.html (unter Überschrift **_conll File Format*), bzw. http://cemantix.org/data/ontonotes.html.

Beispielauszug aus *cctv_0001.v4_gold_conll*:

```
#begin document (bc/cctv/00/cctv_0001); part 000
...
bc/cctv/00/cctv_0001   1    0        Hayao    NNP  (TOP(S(NP(NP*        -    -    -   Speaker#1   (PERSON*        (ARG0*    (ARG0*      -
bc/cctv/00/cctv_0001   1    1         Tada    NNP               *)       -    -    -   Speaker#1           *)           *         *      -
bc/cctv/00/cctv_0001   1    2            ,       ,               *        -    -    -   Speaker#1           *            *         *      -
bc/cctv/00/cctv_0001   1    3    commander      NN         (NP(NP*)       -    -    -   Speaker#1           *            *         *      -
bc/cctv/00/cctv_0001   1    4           of      IN            (PP*        -    -    -   Speaker#1           *            *         *      -
bc/cctv/00/cctv_0001   1    5          the      DT            (NP*        -    -    -   Speaker#1           *            *         *      -
bc/cctv/00/cctv_0001   1    6     Japanese     NNP               *        -    -    -   Speaker#1      (NORP)            *         *      -
bc/cctv/00/cctv_0001   1    7        North     NNP               *        -    -    -   Speaker#1      (ORG*             *         *      -
bc/cctv/00/cctv_0001   1    8        China     NNP               *        -    -    -   Speaker#1           *            *         *      -
bc/cctv/00/cctv_0001   1    9         Area     NNP               *        -    -    -   Speaker#1           *            *         *      -
bc/cctv/00/cctv_0001   1   10         Army     NNP            *))))       -    -    -   Speaker#1           *)           *)        *)     -
...
```

Aufgabe 1. Extrahieren Sie den **Text** (d.h. die sequenzielle Folge der Tokens) aus der Datei *cctv_0001.v4_gold_conll* in eine einfache Textdatei *cctv_0001.v4_gold_conll-text.txt* automatisch mit einem selbstgeschriebenen Programm.
Hinweise:

- In der Ausgabedatei soll ein Satz pro Zeile gespeichert werden.
- Die einzelnen Tokens in einem Satz sollten jeweils durch ein Leerzeichen getrennt werden. Eine Entfernung der Leerzeichen vor den Satzzeichen ist nicht nötig, d.h. man muss die Tokenisierung nicht rückgängig machen.
- Dokumententeilgrenzen in der CoNLL-Datei müssen nicht berücksichtigt werden. (Die Dokumententeile sind in der CoNLL-Datei durch Zeilen mit "#" am Anfang gekennzeichnet.)

[1] http://cemantix.org/data/ontonotes.html

Aufgabe 2. Extrahieren Sie aus der CoNLL-Datei **die Menge der möglichen Label-Werte** für folgende Annotationsebenen mit einem weiteren Skript. Das Ergebnis der Extraktion kann wahlweise auf die Kommandozeile oder in Ausgabedateien erfolgen.

1. **Wortart**[2]
2. **Named Entity**[3]

Aufgabe 3. Extrahieren Sie alle **Named Entities** nach ihren Labels gruppiert und alphabetisch sortiert.

- Hier werden also die Named Entities selbst gesucht, nicht die Labels der Named Entities.

  Z. B.: Für die ersten zwei Wörter im Beispielsatz würde man für Aufgabe 2 den Label *PERSON* extrahieren, und für die aktuelle Aufgabe die Named-Entity-Instanz *Hayao Tada*.

- Je Named-Entity-Label soll eine eigene Ausgabedatei erzeugt werden, die nach dem folgenden Schema benannt werden: `<LABEL>.txt`.
- In jeder Textdatei soll eine Named-Entity-Instanz pro Zeile gespeichert werden.
- Die Instanzen sollten alphabetisch sortiert sein.
- Mehrworteinheiten sind als Solche zu erkennen und zu extrahieren.

  Z. B.: *Hayao Tada* ist **eine** Instanz in der Ausgabedatei *PERSON.txt*.

Abzugeben sind die Implementierungen und die erzeugten Ausgabedateien.

[2] Für Interessierte: Die Annotation verwendet das *Penn-Treebank-Tagset*, beschrieben z.B. hier: http://www.computing.dcu.ie/~acahill/tagset.html.
[3] Für Interessierte: Die annotierten Named Entities werden hier kurz aufgelistet: *OntoNotes Release 5.0.pdf*, S.21f. – https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf

Zusatzaufgabe Laden Sie die Trial-Daten des CoNLL2009-Shared-Tasks für Spanisch herunter: http://ufal.mff.cuni.cz/conll2009-st/trial/CoNLL2009-ST-Spanish-trial.zip. Die enthaltene Datei *CoNLL2009-ST-spanish-trial.txt* dient als Datengrundlage für die Aufgabe. **Exrahieren Sie alle Konstituenten, die von einem verbalen Wurzelknoten des jeweiligen Satzes syntaktisch abhängen.** Betrachten Sie jeweils die **manuellen** Annotationsebenen (siehe Hinweise). Für die Aufgabe kann man zwei Schwierigkeitsgrade implementieren:

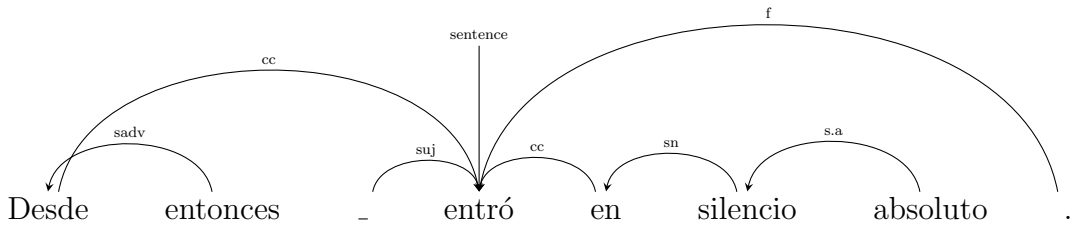
a) Jeweils ausschließlich das **Kopfwort des Dependenten** extrahieren. Das ist das Wort, das direkt vom Prädikat abhängt.

b) Die **ganze abhängige Wortsequenz** extrahieren, die vom dependenten Knoten abgedeckt wird.

Beispielsatz (mit Benennung einzelner Spalten in der ersten Zeile; lange Annotationen wurden übersichtshalber gekürzt dargestellt (...)):

ID	FORM	.	.	POS	.	.	.	HEAD	.	DEPREL	.	.	.	.
1	Desde	desde	desde	s	s	postype=...	postype=...	4	4	cc	cc	_	_	argM-tmp
2	entonces	entonces	entonces	r	r	_	_	1	1	sadv	sadv	_	_	_
3	_	_	_	p	p	_	_	4	4	suj	suj	_	_	arg1-tem
4	entró	entrar	entrar	v	v	postype=...	postype=...	0	0	sentence	sentence	Y	entrar.b2	_
5	en	en	en	s	s	postype=...	postype=...	4	4	cc	cc	_	_	arg2-efi
6	silencio	silencio	silencio	n	n	postype=...	postype=...	5	5	sn	sn	_	_	_
7	absoluto	absoluto	absoluto	a	a	postype=...	postype=...	6	6	s.a	s.a	_	_	_
8	.	.	.	f	f	punct=...	punct=...	4	4	f	f	_	_	_

Die Beispielannotation kodiert diese Dependenzstruktur:



Ausgabeformat für einen Eintrag:

```
<Verbaler Wurzelknoten>
\t <Dependenzrelation> : <Dependent>
```

Ausgabebeispiel für den obigen Satz mit Unteraufgabe a):

```
entró
    cc : Desde
    suj : _
    cc : en
    f : .
```

Ausgabebeispiel für den obigen Satz mit Unteraufgabe b):

```
entró
    cc : Desde entonces
    suj : _
    cc : en silencio absoluto
    f : .
```

Hinweise

- Eine kurze Beschreibung der Spalten finden Sie hier: http://ufal.mff.cuni.cz/conll2009-st/task-description.html#Dataformat. Folgende Spalten sind als Grundlage der Extraktion zu empfehlen (auch im Satzbeispiel gekennzeichnet):
  - Spalte 1: `ID` – Wortnummer im Satz, durchnummeriert ab 1
  - Spalte 2: `FORM` – Wortform
  - Spalte 5: `POS` – manuelle Part-of-speech-Annotation: Für die Aufgabe sollte man nur die verbalen Prädikate als Wurzelknoten beachten, die durch *v* ausgezeichnet sind.
  - Spalte 9: `HEAD` – manuelle Annotation der Dependenstruktur: Die Wortnummer des Mutterknotens im Dependenzbaum wird für das aktuelle Wort angegeben. Der Wurzelknoten wird durch die Nummer 0 ausgezeichnet – das ist in der Regel das verbale Prädikat im Satz.
  - Spalte 11: `DEPREL` – manuelle Annotation der Dependenzrelation. Eine Beschreibung der Dependenzrelationen finden Sie in der Datei *tagsets.pdf* (im heruntergeladenen zip-Ordner).
- Bei Teilaufgabe b) sollte die Ausgabe die abhängigen Konstituenten in der ursprünglichen Reihenfolge der Wörter im Text enthalten (z.B. *Desde entonces* und nicht *entonces Desde*).