Annotation Formats: Types and examples

COMP61332: Text Mining
Week 1
Riza Batista-Navarro

Types of annotation formats

Boundary notation

Inline markup language elements

Stand-off

- delimiter-separated values (DSV)
- JSON

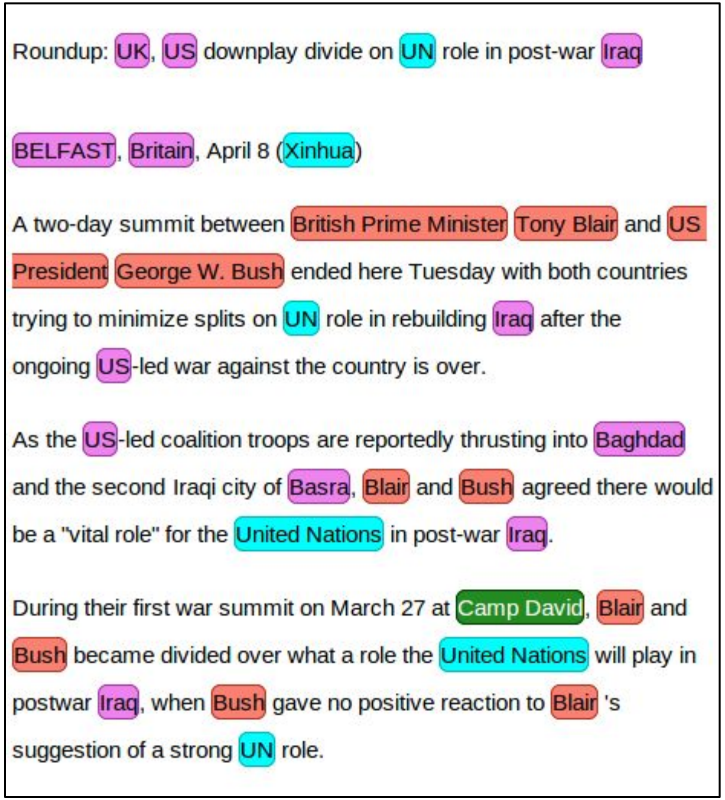
Boundary Notation

Done at the level of individual tokens

How do we encode units of interest spanning several tokens?

BIO: B=Begin, I=Inside, O=Outside

Named entity tags in BIO



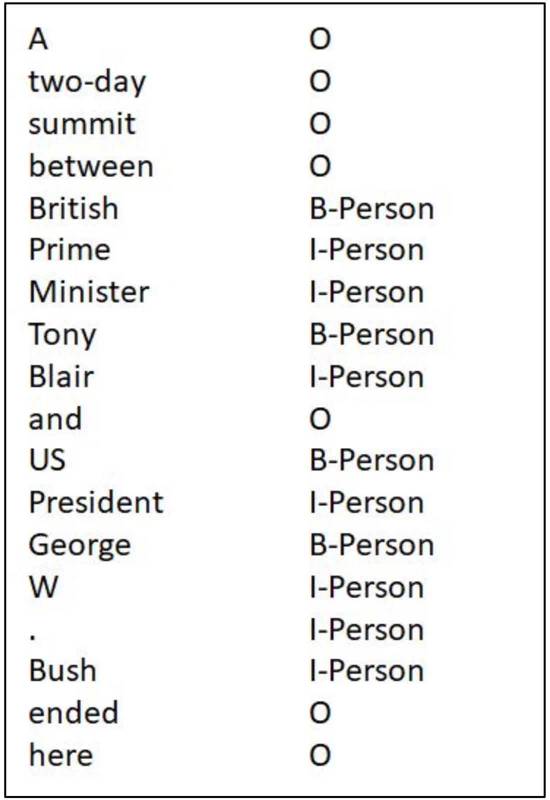
Roundup: UK, US downplay divide on UN role in post-war Iraq

BELFAST, Britain, April 8 (Xinhua)

A two-day summit between British Prime Minister Tony Blair and US President George W. Bush ended here Tuesday with both countries trying to minimize splits on UN role in rebuilding Iraq after the ongoing US-led war against the country is over.

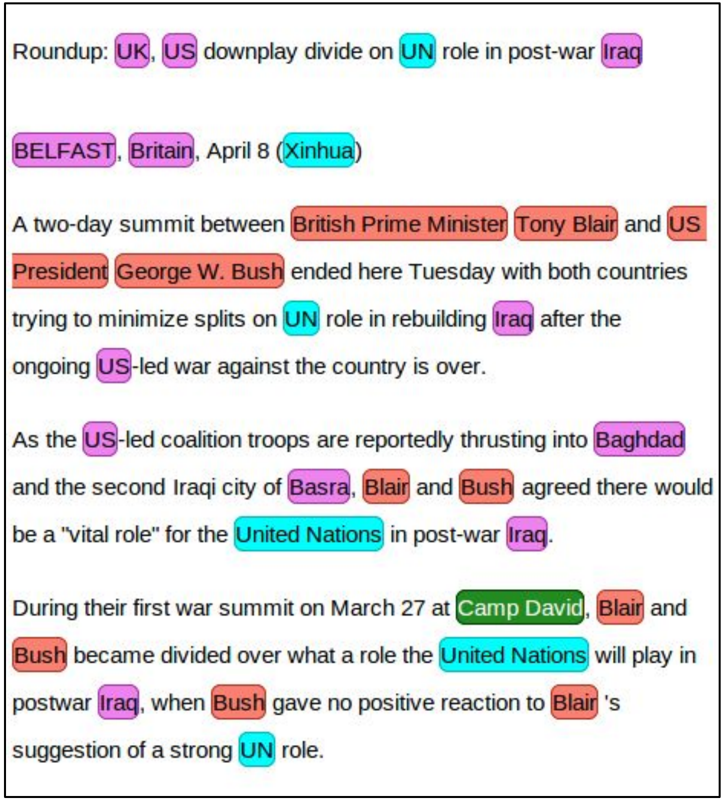
As the US-led coalition troops are reportedly thrusting into Baghdad and the second Iraqi city of Basra, Blair and Bush agreed there would be a "vital role" for the United Nations in post-war Iraq.

During their first war summit on March 27 at Camp David, Blair and Bush became divided over what a role the United Nations will play in postwar Iraq, when Bush gave no positive reaction to Blair 's suggestion of a strong UN role.



A	O
two-day	O
summit	O
between	O
British	B-Person
Prime	I-Person
Minister	I-Person
Tony	B-Person
Blair	I-Person
and	O
US	B-Person
President	I-Person
George	B-Person
W	I-Person
.	I-Person
Bush	I-Person
ended	O
here	O

Named entity tags in BIO



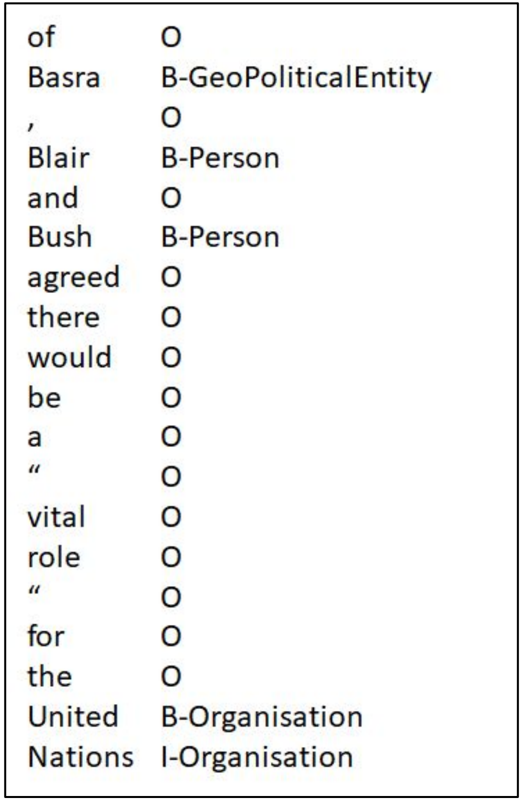
Roundup: UK, US downplay divide on UN role in post-war Iraq

BELFAST, Britain, April 8 (Xinhua)

A two-day summit between British Prime Minister Tony Blair and US President George W. Bush ended here Tuesday with both countries trying to minimize splits on UN role in rebuilding Iraq after the ongoing US-led war against the country is over.

As the US-led coalition troops are reportedly thrusting into Baghdad and the second Iraqi city of Basra, Blair and Bush agreed there would be a "vital role" for the United Nations in post-war Iraq.

During their first war summit on March 27 at Camp David, Blair and Bush became divided over what a role the United Nations will play in postwar Iraq, when Bush gave no positive reaction to Blair 's suggestion of a strong UN role.



of	O
Basra	B-GeoPoliticalEntity
,	O
Blair	B-Person
and	O
Bush	B-Person
agreed	O
there	O
would	O
be	O
a	O
“	O
vital	O
role	O
“	O
for	O
the	O
United	B-Organisation
Nations	I-Organisation

Boundary Notation

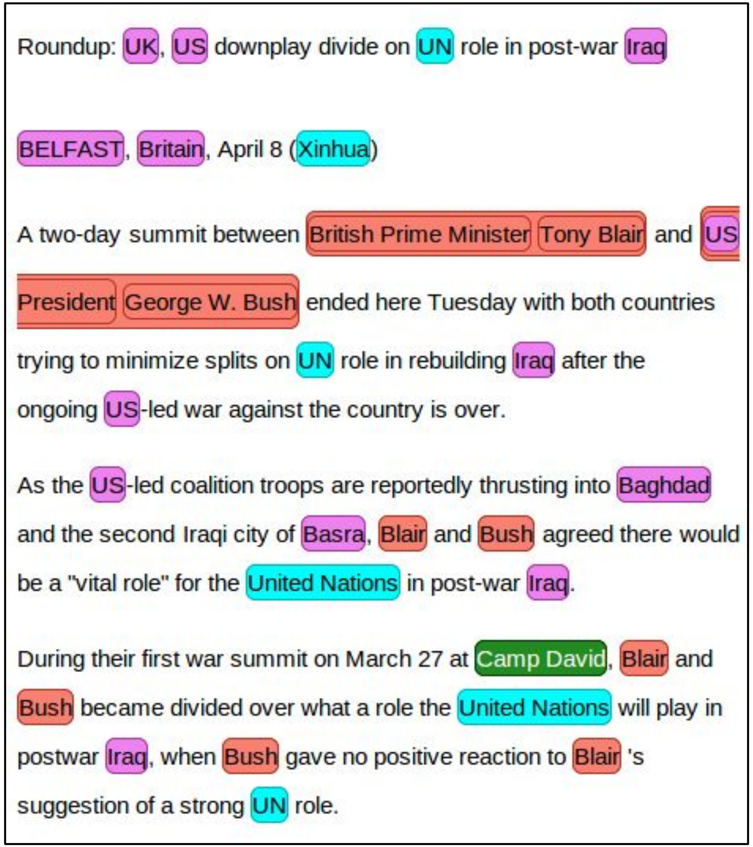
Strengths

- simple

Limitations

- cannot handle hierarchical or structured annotations, e. g., nested entities, relations, events

Example: Nested entities (NEs)



Roundup: UK, US downplay divide on UN role in post-war Iraq

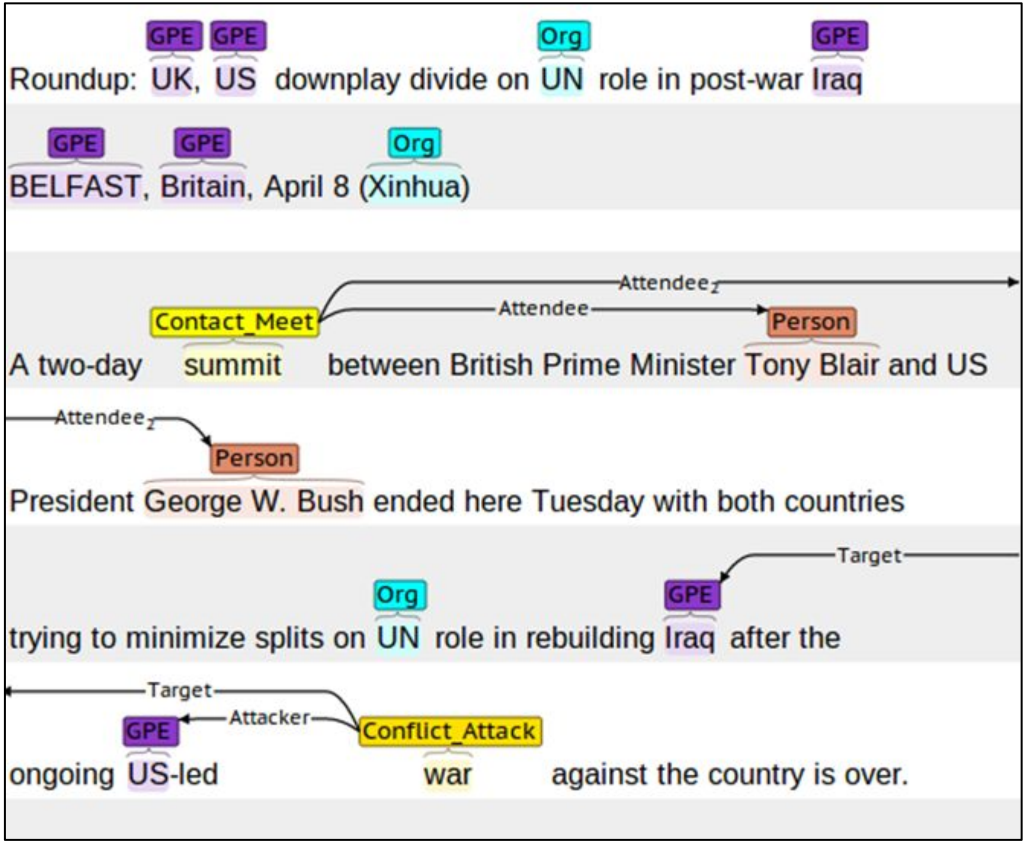
BELFAST, Britain, April 8 (Xinhua)

A two-day summit between British Prime Minister Tony Blair and US President George W. Bush ended here Tuesday with both countries trying to minimize splits on UN role in rebuilding Iraq after the ongoing US-led war against the country is over.

As the US-led coalition troops are reportedly thrusting into Baghdad and the second Iraqi city of Basra, Blair and Bush agreed there would be a "vital role" for the United Nations in post-war Iraq.

During their first war summit on March 27 at Camp David, Blair and Bush became divided over what a role the United Nations will play in postwar Iraq, when Bush gave no positive reaction to Blair 's suggestion of a strong UN role.

Example: Events



Roundup: UK, US downplay divide on UN role in post-war Iraq

BELFAST, Britain, April 8 (Xinhua)

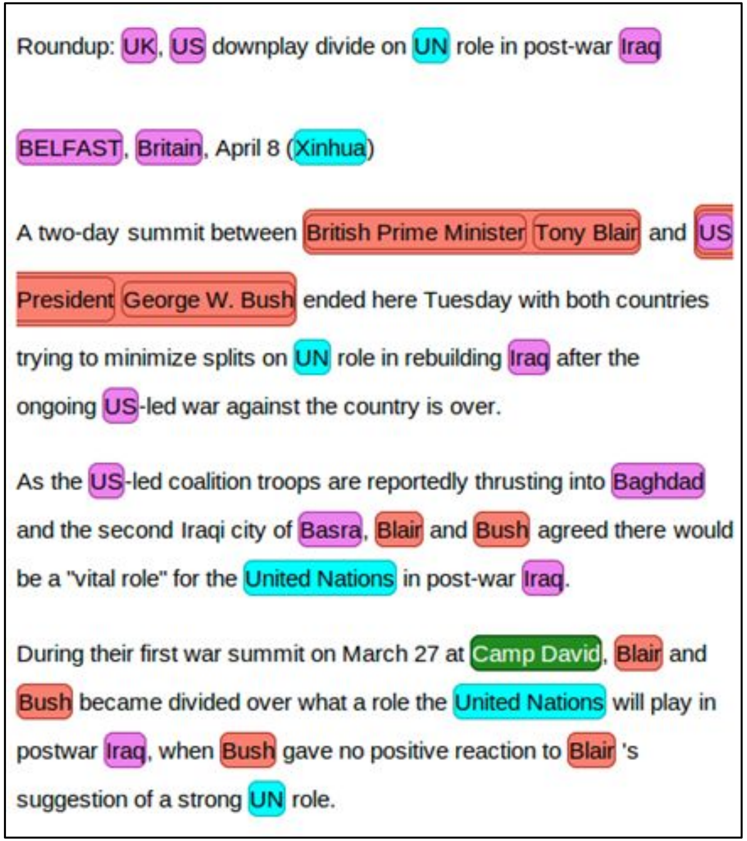
A two-day summit between British Prime Minister Tony Blair and US President George W. Bush ended here Tuesday with both countries trying to minimize splits on UN role in rebuilding Iraq after the ongoing US-led war against the country is over.

Inline markup language elements

By addition of markup tags within text

e.g., **HTML**, **XML**

Nested NEs as inline XML elements



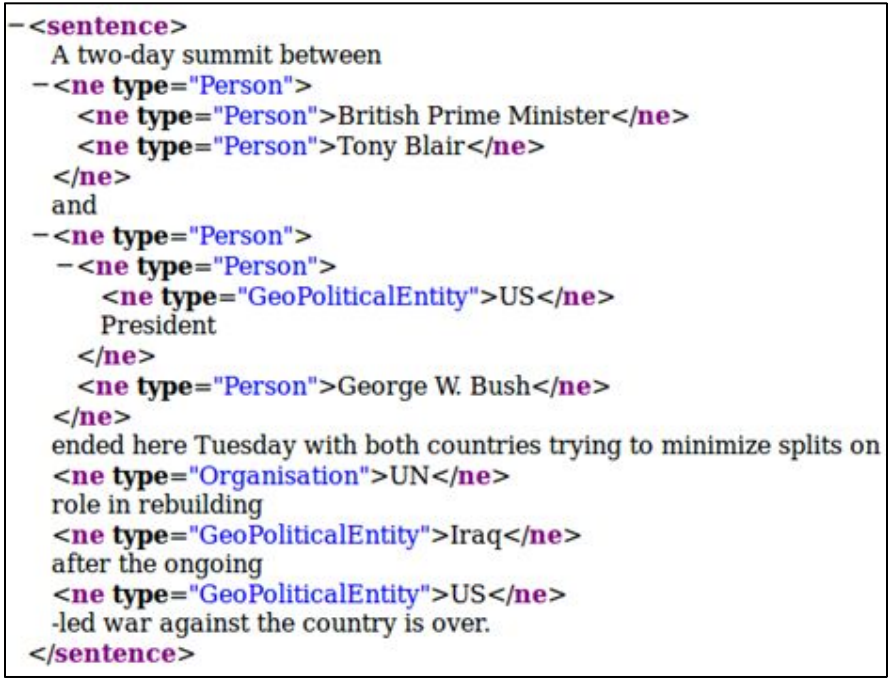
Roundup: UK, US downplay divide on UN role in post-war Iraq

BELFAST, Britain, April 8 (Xinhua)

A two-day summit between British Prime Minister Tony Blair and US President George W. Bush ended here Tuesday with both countries trying to minimize splits on UN role in rebuilding Iraq after the ongoing US-led war against the country is over.

As the US-led coalition troops are reportedly thrusting into Baghdad and the second Iraqi city of Basra, Blair and Bush agreed there would be a "vital role" for the United Nations in post-war Iraq.

During their first war summit on March 27 at Camp David, Blair and Bush became divided over what a role the United Nations will play in postwar Iraq, when Bush gave no positive reaction to Blair 's suggestion of a strong UN role.



```
-<sentence>
   A two-day summit between
 -<ne type="Person">
     <ne type="Person">British Prime Minister</ne>
     <ne type="Person">Tony Blair</ne>
   </ne>
   and
 -<ne type="Person">
   -<ne type="Person">
       <ne type="GeoPoliticalEntity">US</ne>
       President
     </ne>
     <ne type="Person">George W. Bush</ne>
   </ne>
   ended here Tuesday with both countries trying to minimize splits on
   <ne type="Organisation">UN</ne>
   role in rebuilding
   <ne type="GeoPoliticalEntity">Iraq</ne>
   after the ongoing
   <ne type="GeoPoliticalEntity">US</ne>
   -led war against the country is over.
 </sentence>
```

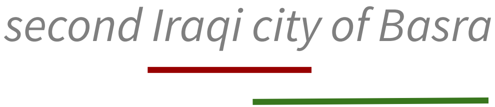
Inline markup language elements

Strengths

- can handle annotations which are hierarchical (e.g., nested NEs, trees) and structured (e.g., events)

Limitations

- requires substantial processing with standard XML parsers
- impossible to encode overlapping/intersecting annotations, e.g.,



second Iraqi city of Basra

Stand-off annotations

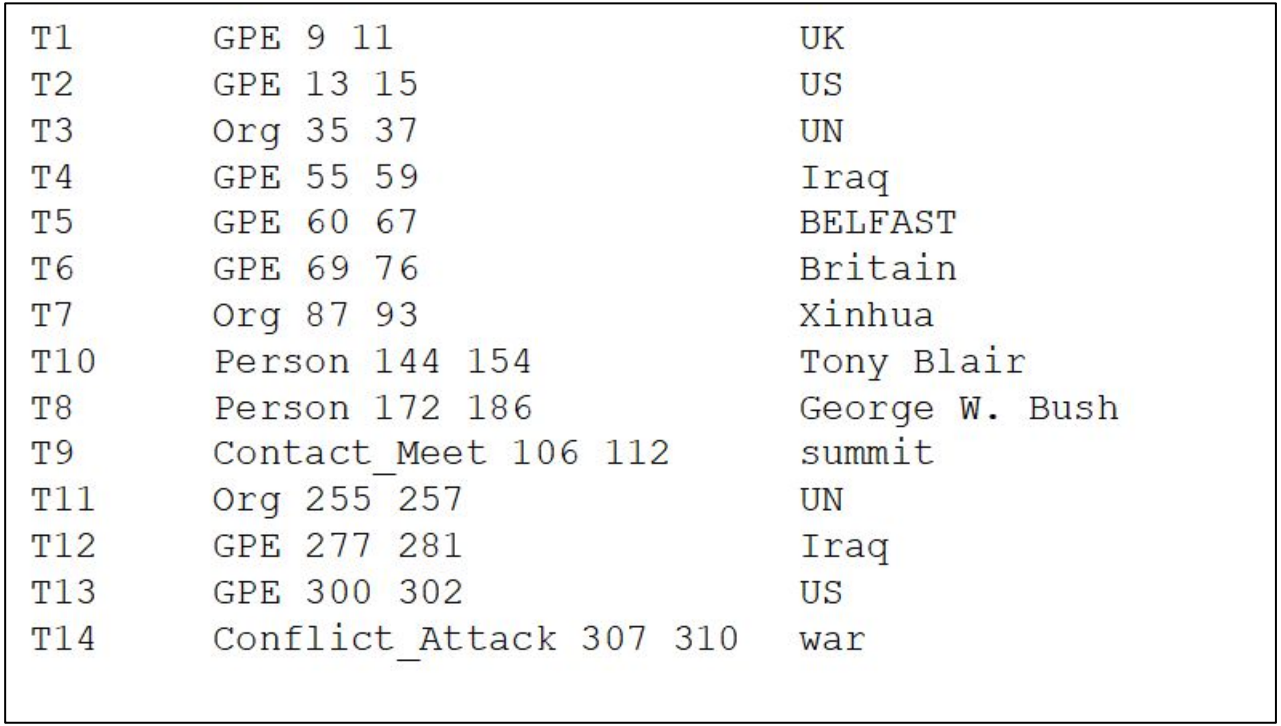
annotations are **stored separately**

requires **a way to link between annotations and text**

links annotations to text using indexing based on **character offsets** (computed over raw text)

Named entities and events in stand-off DSV

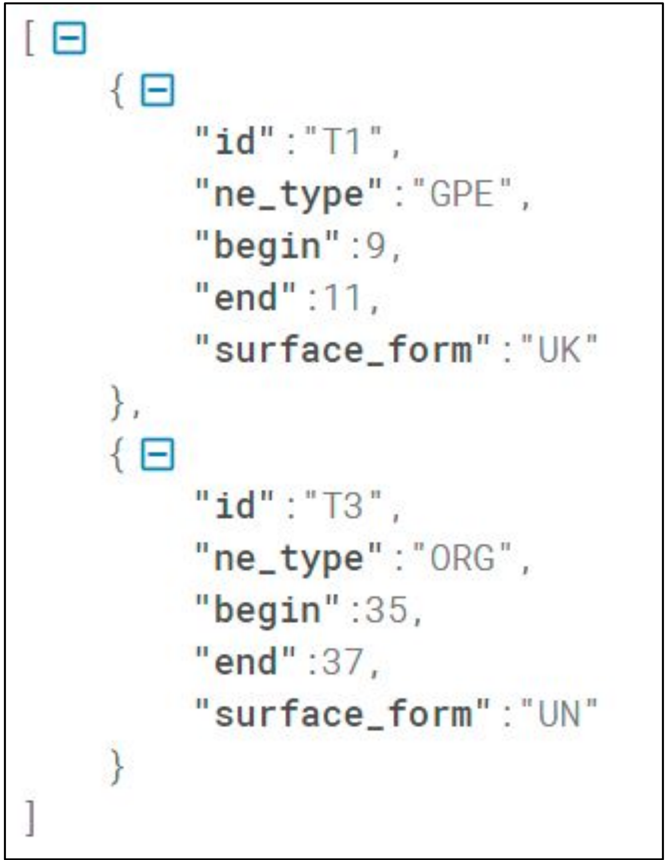
Delimiter-separated values (DSV)



```
T1      GPE 9 11                    UK
T2      GPE 13 15                   US
T3      Org 35 37                   UN
T4      GPE 55 59                   Iraq
T5      GPE 60 67                   BELFAST
T6      GPE 69 76                   Britain
T7      Org 87 93                   Xinhua
T10     Person 144 154              Tony Blair
T8      Person 172 186              George W. Bush
T9      Contact_Meet 106 112        summit
T11     Org 255 257                 UN
T12     GPE 277 281                 Iraq
T13     GPE 300 302                 US
T14     Conflict_Attack 307 310     war
```

Named entities and events in stand-off JSON

JavaScript Object Notation (JSON)



```
[ ⊟
    { ⊟
        "id":"T1",
        "ne_type":"GPE",
        "begin":9,
        "end":11,
        "surface_form":"UK"
    },
    { ⊟
        "id":"T3",
        "ne_type":"ORG",
        "begin":35,
        "end":37,
        "surface_form":"UN"
    }
]
```

Stand-off annotations

Strengths

- original raw text is left untouched
- can handle structured and overlapping annotations

Limitations

- not readily human-readable