Edinburgh-LTG: TempEval-2 System Description

Claire Grover, Richard Tobin, Beatrice Alex and Kate Byrne
University of Edinburgh
Edinburgh, United Kingdom
`{grover, richard, balex, kbyrne3}@inf.ed.ac.uk`

Abstract

We describe the Edinburgh information extraction system which we are currently adapting for analysis of newspaper text as part of the SYNC3 project. Our most recent focus is geospatial and temporal grounding of entities and it has been useful to participate in TempEval-2 to measure the performance of our system and to guide further development. We took part in Tasks A and B for English.

1 Background

The Language Technology Group (LTG) at Edinburgh has been active in the field of information extraction (IE) for a number of years. Up until recently our main focus has been in biomedical IE (Alex et al., 2008) but we have also been pursuing projects in other domains, e.g. digitised historical documents (Grover et al., 2010) and we are currently participants in the EU-funded SYNC3 project where our role is to analyse news articles and establish spatio-temporal and other relations between news events. As a step towards this goal, we have been extending and adapting our IE pipeline to ground spatial and temporal entities. We have developed the Edinburgh Geoparser for georeferencing documents and have evaluated our system against the SpatialML corpus, as reported in Tobin et al. (2010). We are currently in the process of developing a rule-based date and time grounding component and it is this component that we used for Task A, which requires systems to identify the extents of temporal named entities and provide their interpretation. The TempEval-2 data also contains event entities and we have adapted the output of our in-house chunker (Grover and Tobin, 2006) to identify events for Task B, which requires systems to identify event denoting words and to compute a range of attributes for them. In future work we will adapt our machine-learning-based relation extrac-

tion component (Haddow, 2008) to recognise relations between spatial and temporal entities and event entities along the lines of the linking tasks.

2 The Edinburgh IE System

Our IE system is a modular pipeline system built around the LT-XML2[1] and LT-TTT2[2] toolsets. Documents are converted into our internal document format and are then passed through a sequence of linguistic components which each add XML mark-up. Early stages identify paragraphs, sentences and tokens. Part-of-speech (POS) tagging is done using the C&C tagger (Curran and Clark, 2003a) and lemmatisation is done using morpha (Minnen et al., 2000).

We use both rule-based and machine-learning named entity recognition (NER) components, the former implemented using LT-TTT2 and the latter using the C&C maximum entropy NER tagger (Curran and Clark, 2003b). We are experimenting to find the best combination of the two different NER views but this is not an issue in the case of date and time entities since we have taken the decision to use the rule-based output for these. The main motivation for this decision arises from the need to ground (provide temporal values for) these entities and the rules for the grounding are most naturally implemented as an elaboration of the rules for recognition.

Our IE pipeline also uses the LT-TTT2 chunker to provide a very shallow syntactic analysis. Figure 1 shows an example of the results of processing at the point where the rule-based NER and chunker have both applied. As can be seen from Figure 1, a positive feature for TempEval-2 is that the verb group analysis provides information about tense, aspect, voice, modality and polarity which translate relatively straightforwardly into the Task B attributes. The noun group analysis provides verbal stem information (e.g.

---

[1] `www.ltg.ed.ac.uk/software/ltxml2`
[2] `www.ltg.ed.ac.uk/software/lt-ttt2`

```
<s id="s1">
 <ng>
  <w p="DT" id="w13">The</w>
  <w p="NN" id="w17" l="announcement" vstem="announce" headn="yes">announcement</w>
 </ng>
 <vg tense="pres" voice="pass" asp="simple" modal="yes" neg="yes">
  <w p="MD" id="w30" pws="yes" l="must" neg="yes">must</w>
  <w p="RB" id="w35" pws="yes" neg="yes">not</w>
  <w p="VB" id="w39" pws="yes" l="be">be</w>
  <w p="VBN" id="w42" pws="yes" l="make" headv="yes">made</w>
 </vg>
 <ng>
  <timex unit="day" trel="same" type="date" id="rb1">
   <w unit="day" trel="same" p="NN" id="w47" l="today">today</w>
  </timex>
 </ng>
 <w p="." id="w52" sb="true">.</w>
</s>
```

Figure 1: Example of NER tagger and chunker output for the sentence "The announcement must not be made today."

`vstem="announce"`) about nominalisations.

Various attributes are computed for `<timex>` elements and these are used by a temporal resolution component to provide a grounding for them. The final output of the IE pipeline contains entity mark-up in "standoff" format where the entities point at the word elements using ids. The date and event entities for "made" and "today" are as follows:

```
<ent tense="pres" voice="pass" neg="yes"
    modal="yes" asp="simple" id="ev1"
    subtype="make" type="event">
  <parts>
    <part ew="w39" sw="w39">made</part>
  </parts>
</ent>

<ent wdaynum="5" day="Friday" date="16"
    month="4" year="2010" unit="day"
    day-number="733877" trel="same"
    type="date" id="rb1">
  <parts>
    <part ew="w47" sw="w47">today</part>
  </parts>
</ent>
```

The date entity has been grounded with respect to the date of writing (16th April 2010). To do the grounding we calculate a `day-number` value for each date where the day number count starts from 1st January 1 AD. Using this unique day number we are able to calculate the date for any given day number as well as the day of the week. We use the day number to perform simple arithmetic to ground date expressions such as "last Monday", "the day after tomorrow" etc. Grounding information is spread across the attributes for `day`, `date`, `month` and `year`. A fully grounded date has a value for all of these while an underspecified date, e.g. "2009", "March 13th", "next year", etc., only has values for some of these attributes.

3 Adaptations for TempEval-2

Our system has been developed independently of TimeML or TempEval-2 and there is therefore a gap between what our system outputs and what is contained in the TempEval-2 data. In order to run our system over the data we needed to convert it into our XML input format while preserving the tokenisation decisions from the original. Certain tokenisation mismatches required that we extend various rules to allow for alternative token boundaries, for example, we tokenise "wasn't" as was + n't whereas the TempEval-2 data contains was + n + 't or occasionally wasn + 't.

Other adaptations fall broadly into two classes: extension of our system to cover entities in TempEval-2 that we didn't previously recognise, and mapping of our output to fit TempEval-2 requirements.

3.1 Extensions

The date and time entities that our system recognises are more like the MUC7 TIMEX entities (Chinchor, 1998) than TIMEX3 ones. In particular, we have focused on dates which can either be fully grounded or which, though underspecified, can be grounded to a precise range, e.g. "last month" can be grounded to a particular month and year given a document creation date and it can be precisely specified if we take it to express a range from the first to last days of the month. TIMEX3 entities can be vaguer than this, for example, entities of type DURATION such as "twenty years", "some time", etc. can be recognised as denoting a temporal period but cannot easily be grounded. To align our output more closely to TempEval-2, we added NER rules to recognise examples

such as "a long time", "recent years", "the past", "years", "some weeks", "10 minutes". In addition we needed to compute appropriate information to allow us to create TempEval-2 values such as `P1W` (period of 1 week).

For event recognition, our initial system created an event entity for every head verb and for every head noun which was a nominalisation. This simple approach goes a long way towards capturing the TempEval-2 events but results in too many false positives and false negatives for nouns. In addition our system did not calculate the information needed to compute the TempEval-2 class attribute. To help improve performance we added attributes to potential event entities based on lookup in lexicons compiled from the training data and from WordNet (Fellbaum, 1998). These attributes contribute to the decision as to whether a noun or verb chunk head should be an event entity or not[3]. The lexicons derived from the training data contain the stems of all the nouns which acted more than once as events as well as information about those predicates which occurred more than once as class ASPECTUAL, I_STATE, REPORTING or STATE in the training data. Where lookup succeeds for event, if class look-up also succeeds then the class attribute is set accordingly. If class look-up fails, the default, OCCURRENCE, is used. The WordNet derived lexicon contains information about whether the first sense of a noun has event or state as a hypernym. As a result of the lexical look-up stage, the noun "work", for example, is marked as having occurred in the training data as an event and as having event as a hypernym for its first sense. The conjunction of these cause it to be considered to be an event entity. For verbs, the only substantive change in our system was to not consider as events all main verb uses of "be" (be happy), "have" (have a meal) and "do" (do the dishes).

3.2 Mapping

For both timex and event entities the creation of the extents files was a straightforward mapping. For the creation of the attributes files, on the other hand, we used stylesheets to construct appropriate values for the TempEval-2 attributes based on the attributes in our output XML. The construction of event attributes is not overly complex: for example, where an event entity is specified as `tense="nonfin"` and

---

[3]Our system does not recognise adjective events. However, passive participles, which are sometimes treated as adjectives in TempEval-2, are frequently treated as verbs in our system and are therefore recognised.

`voice="pass"` the TempEval-2 tense attribute is given the value PASTPART. For modality our attribute only records whether a modal verb is present or not, so it was necessary to set the TempEval-2 modality attribute to the actual modal verb inside the verb group.

For timex entities, a single value for the value attribute had to be constructed from the values of a set of attributes on our entity. For example, the information in `date="16", month="4" year="2010"` has to be converted to 2010-04-16. For durations other attributes provide the relevant information, for example for "two days" the attributes `unit="day", quty="2"` are used to create the value P2D (period of 2 days).

4 Evaluation and Error Analysis

The recognition results for both timex and event extents are shown in Table 1. For Task A (timex) we achieved a close balance between precision and recall, while for Task B (events) we erred towards recall at some cost to precision.

Task	Precision	Recall	F1
Task A	0.85	0.82	0.84
Task B	0.75	0.85	0.80

Table 1: Extent Results

For timex entities our false negatives were all entities of the vaguest kind, for example, "10-hour", "currently", "third-quarter", "overnight", "the week": these are ones which the original system did not recognise and for which we added extra rules, though evidently we were not thorough enough. The false positives were mostly of the kind that would usually be a date entity but which were not considered to be so in the key, for example, "1969", "Oct 25", "now", "the past", "a few days". In two cases the system mistakenly identified numbers as times ("1.02", "2.41").

For event entities we had 73 false negatives. Some of these were caused by verbs being mistagged as nouns ("complies", "stretch", "suit") while others were nouns which didn't occur in the WordNet derived lexicon as events. There were 143 event false positives. Some of these are clearly wrong, for example, "destruction" in "weapons of mass destruction" while others are a consequence of the subtle distinctions that the TempEval-2 guidelines make and which our shallow approach cannot easily mimic.

Table 2 shows the results for attribute detection for both tasks. In the case of timex attributes

Task	Attribute	Score
Task A	type	0.84
	value	0.63
Task B	polarity	0.99
	pos	0.97
	modality	0.99
	tense	0.92
	aspect	0.98
	class	0.76

Table 2: Attribute Results

there was a set of entities which had systematically wrong values for both type and value: these were dates such as "this week" and "last week". These should have had DATE as their type and a value such as 1998-W19 to indicate exactly which week in which year they denote. Our date grounding does not currently cover the numbering of weeks in a year and so it would not have been possible to create appropriate values. Instead we incorrectly treated these entities as being of type DURATION with value P1W. Many of the remaining errors were value errors where the system resolved relative dates as past references when they should have been future or vice versa. For example, the value for "Monday" in "He and Palestinian leader Yasser Arafat meet separately Monday with ..." should have been 1998-05-04 but our system interpreted it as the past Monday, 1998-04-27. There were a few cases where the value was correct but insufficient, for example for "a year ago" the system returned 1988 when it should have produced 1988-Q3.

Our scores for event attributes were high for all attributes except for class. The high scoring attributes were derived from the output of our chunker and demonstrate the quality of this component. There does not appear to be a particular pattern behind the small number of errors for these attributes except that errors for the pos attribute reflect POS tagger errors and there were some combined tense and modality errors where "will" and "would" should have been interpreted as future tense but were instead treated as modals. The class attribute represents information that our system had not previously been designed to determine. We computed the class attribute in a relatively minimal way. Since the class value is OCCURRENCE in nearly 60% of events in the training data, we use this as the default but, as described in Section 3, we override this for events which are in our training data-derived lexicon as REPORTING, ASPECTUAL, I_STATE or STATE. We do not at-

tempt to assign the I_ACTION class value and nearly half of our class errors result from this. Another set of errors comes from missing REPORTING events such as "alleging", "telegraphed" and "acknowledged".

Acknowledgements

References

Beatrice Alex, Claire Grover, Barry Haddow, Mijail Kabadjov, Ewan Klein, Michael Matthews, Richard Tobin, and Xinglong Wang. 2008. Automating curation using a natural language processing pipeline. *Genome Biology*, 9(Suppl 2).

Nancy A. Chinchor. 1998. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Fairfax, Virginia.

James R. Curran and Stephen Clark. 2003a. Investigating GIS and smoothing for maximum entropy taggers. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-03)*, pages 91–98. Budapest, Hungary.

James R. Curran and Stephen Clark. 2003b. Language independent NER using a maximum entropy tagger. In *Proceedings of the 7th Conference on Natural Language Learning*, Edmonton, Alberta, Canada.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Claire Grover and Richard Tobin. 2006. Rule-based chunking and reusability. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.

Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball. 2010. Use of the Edinburgh geoparser for georeferencing digitised historical collections. *Phil. Trans. R. Soc. A*.

Barry Haddow. 2008. Using automated feature optimisation to create an adaptable relation extraction system. In *Proc. of BioNLP 2008*, Columbus, Ohio.

Guido Minnen, John Carroll, and Darren Pearce. 2000. Robust, applied morphological generation. In *Proceedings of the 1st International Natural Language Generation Conference*, Mitzpe Ramon, Israel.

Richard Tobin, Claire Grover, Kate Byrne, James Reid, and Jo Walsh. 2010. Evaluation of georeferencing. In *Proceedings of Workshop on Geographic Information Retrieval (GIR'10)*.

---

[4]http://www.sync3.eu/