Massively Increasing TIMEX3 Resources: A Transduction Approach

The University Of Sheffield.

Universitat d'Alacant
Universidad de Alicante

Leon Derczynski (leon@dcs.shef.ac.uk), Hector Llorens (hllorens@dlsi.ua.es) and Estela Saquete (stela@dlsi.ua.es)

Introduction

Being able to identify **times** is important to practical NLP:

- Question answering
- Information retrieval
- Forensics – disaster analysis
- Summarisation
- Time-series analysis

Temporal expressions (**timexes**) are a basic part of time in language. They refer to a period, a specific time, or a recurring point.

Time expressions:

```
I enjoyed March 17th
Let's go to Istanbul for a week
My date of birth is 13.07.1974
Every Saturday, I play Scrabble
```

How can we spot these timexes in text?

- Rule-based systems
- Machine-learning based systems

Rule-based systems are **fragile**, but can perform well on common expression types

Statistical systems perform well, and cope with previoously-unseen timex formulations. However, they **require data**; the more the better.

Where do we get timexes from?

Human-annotated timex resources have developed over the past ten years. These typically include gold-standard human annotations showing where timexes are, and what the interpreted value of that timex should be.

*One **day** in March → `XXXX-03-XX`*

These are typically in one of two formats:

TIMEX2

Old standard
Long phrases marked as one expression
Lots of data

```
The Yankees had just finished <TIMEX2 val="1998-10-02TEV">a
draining evening</TIMEX2> with a 4-0 decision over the Rangers
```

TIMEX3

Newer standard
Short phrases
Not much data
Includes events and other temporal information in text

```
until <TIMEX3 tid="t31" type="DURATION" value="P90D"
temporalFunction="false" functionInDocument="NONE">90
days</TIMEX3>  later
```

TIMEX2 Corpora

Resource name	Tokens	Timexes
WikiWars	120K	2 681
ACE 2004 TERN	54.6K	8 047
ACE 2005	260K	5 483
TIDES dialogue	31.6K	3 541
Total	**466K**	**19 752**

TIMEX3 Corpora

TimeBank v1.2	68.5K	1 414
AQUAINT	34.1K	609
TempEval-2 test	5.5K	81
TimenEval	7.9K	214
Total	**116K**	**3 289**

Creating more TIMEX3

Because the standards are very clear and mapping can be made programmatically between them in many cases, **basic automatic transduction** is a simple way to start.

The T2T3 tool processes TIMEX2 in a constrained environment and transduces them to TIMEX3, based on the standards.

T2T3 ⇒

Input: TIMEX2-compliant document; output: TIMEX3 entities.

How well does the tool perform?

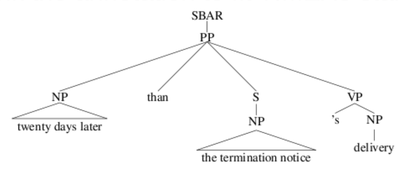
Only works on **clean** TIMEX2
Only works on **latin charset**
Cannot break up long expressions into multiple TIMEX3 annotations
Fails on **embedded expressions**.

What techniques will improve this?

Breaking apart long timexes

TIMEX3 prefers to annotate the **shortest phrase possible**, and is very granular. **TIMEX2** annotations are **comprehensive**, and can include **multiple sub-timexes**.

Our improved tool uses a **model of timex length** and **constituent parsing** to identify the most-relevant part of a long timex and shrink the annotation to fit TIMEX3 standards



Event-based expressions

These consist of a timex and an **event**, often linked together by a **temporal signal** – a special temporal co-ordinator.

We find the signal, **break apart the phrase** based on a parse tree, find timex and event, and **create a TLINKed annotation**

*The Tuesday **after** the party*
["The Tuesday", signal."after", "the party"]
*["The", timex3."**Tuesday**", signal."after", "the", event."**party**"]*
`+ <TLINK timeID="timex3" relatedToEventInstance="event" signalID="signal" relType="" />`

Techincal improvements

Nested expressions

T2T3 resolves nesting by **isolating the atomic TIMEX2 annotations** and **adding TimeML signal annotations** between them.

before <the week of <the seventh> until <the eleventh> >

Character set handling

Added support for **extra encodings** – particularly useful for WikiWars corpus which includes named entities in local text (e.g. Korean, Cyrillic). Uses the Mozilla `chardet` library.

```
known in mainland China as the "<TIMEX2 val="P10Y">Ten Year</TIMEX2>'s
Civil War" (simplified Chinese: 十年内战 ; pinyin: Shínián Nèizhàn)
```

Technical improvements

- Migrated to **NLTK** for linguistic processing
- **PoS cache** for ~70x speedup
- Cut out lemmatisation
- Source code **publicly available**

Resultant resources

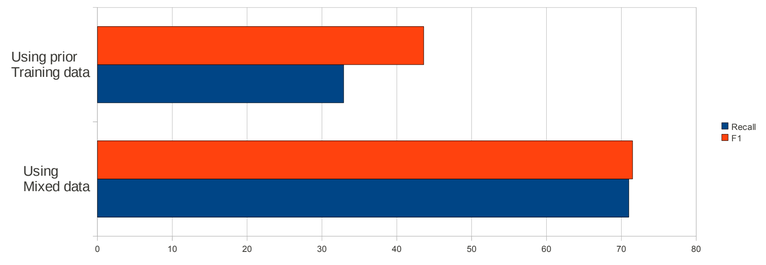
Expanded the body of available TIMEX3 data

Corpus	Date	Time	Dur	Set	Total
Wikiwars	2 323	230	93	30	2 676
ACE 2004	4 697	947	1 759	235	7 638
ACE 2005	3 024	628	1 406	141	5 199
TIDES dialogue	1 684	141	1 402	63	3 290
Total	**14 492**	**2 394**	**4 835**	**523**	**18 803**

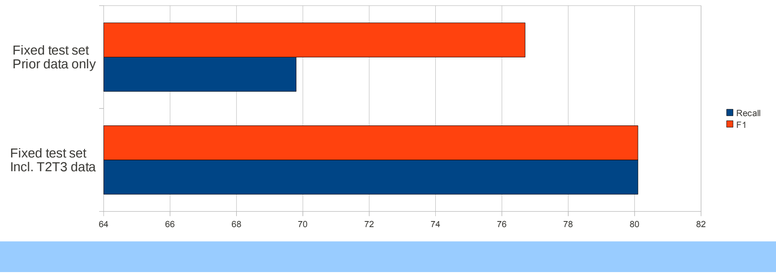
Previously, there were only **3 289** TIMEX3s; now we can use **22 092**.

Evaluation

1. How difficult is new data is to recognise using prior models?
2. How does the tagger behave with the new data?



3. As a baseline, what is performance against a held-out test set of prior data, trained on the rest of the prior data?
4. With the same test set as in 3., what effect does adding the new T2T3 data to the training set have?



Conclusion

We have updated a prototype tool to create a robust means of converting TIMEX2 resources to TimeML / TIMEX3.

Using the new version, we have mapped many older corpora into the new format, **multiplying the amount of TIMEX3 data available by almost seven**.

Use of the increased volume of training data yielded **instant improvements over the state of the art** in recognition.

Further, the new datasets are **very diverse** and cannot be recognised using existing models..

Further work

T2T3 can be extended to other languages, to process TIMEX2 data in **Spanish and German**.

Further, post-hoc **validation** makes the dataset easy to re-use in other experiments and systems.

Finally, the varied forms of expression are a good resource for building **temporal normalisation** systems.

References: Llorens, H et al. (2010). "*TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2*"
Saquete, E and Pustejovsky, J (2011). "*Automatic transformation from TIDES to TimeML annotation*"
Derczynski, L and Gaizauskas, R (2010). *"Analysing temporally annotated corpora with CAVaT"*
Mazur, P and Dale, R (2010). *"Wikiwars: A new corpus for research on temporal expressions"*
Mani, I et al. (2001). *"Guidelines for annotating temporal information"*
Pustejovsky, J et al. (2003). *"TimeML: Robust specification of event and temporal expressions in text"*