Annotating the Enron Email Corpus with Number Senses

Stuart Moore*, Sabine Buchholz†, Anna Korhonen*

* Computer Laboratory, University of Cambridge
† Toshiba Research Europe, Cambridge
stuart.moore@cl.cam.ac.uk, sabine.buchholz@crl.toshiba.co.uk, anna.korhonen@cl.cam.ac.uk

Abstract

The Enron Email Corpus provides "Real World" text in the business email domain, which is a target domain for many speech and language applications. We present a section of this corpus annotated with number senses - labelling each number as a date, time, year, telephone number etc. We show that sense categories and their frequencies are very different in this domain than in newswire text. The annotated corpus can provide valuable material for the development of number sense disambiguation techniques. We have released the annotations into the public domain, to allow other researchers to perform comparisons.

1. Introduction

Digits within text can be used to express a variety of different meanings. For example, *'2008'* could mean a year (*'Spain won the 2008 European Championships'*), a time (*'Our train arrives at 2008'*), a quantity (*'There were 2008 sweets in the jar'*) or a telephone number (*'Call 0800 508 2008 for more details'*). Pronunciation can vary according to the context - e.g. *'1990'* would be pronounced *'nineteen ninety'* as a year, but *'one thousand, nine hundred and ninety'* as a quantity.

Although most systems dealing with text will have some methods for dealing with numbers, Numbers Sense Disambiguation techniques could enhance performance of various NLP tasks, e.g.

- **Speech Synthesis** - Different number senses have different pronunciations.
- **Information Extraction** - identifying documents that refer to the year 2000 instead of 2000 as a quantity.
- **Machine Translation** - Different number senses may require different translations.

To enable effective development and wider applications of Number Sense Disambiguation systems, such as the ones proposed by Yarowsky (1996), Sproat et al. (2001) and Moore et al. (2009), text in a variety of domains needs to be annotated. However the only publicly available corpus that we are aware of is that of Sproat et al. (2001). They annotated text from four sources - the largest is a subsection of the North American News Text corpus (Graff, 1995) (newswire text). Sproat et al. were focusing on text normalisation of acronyms and abbreviations; numbers were labelled with a sense, but this was not their main focus.

It is likely that number senses (like word senses) vary between different domains and text types (e.g. sports, biomedical and legal text may exhibit different sense types and distributions). To investigate this, we are annotating another important domain (business email) where the senses and their context are likely to be different from those found in newswire text.

Our annotations are released into the public domain, and are available from `http://www.cl.cam.ac.uk/~stjm2/enron/` along with some brief guidelines on usage.

2. Corpus

The Enron Email corpus (Klimt and Yang, 2004) consists of emails that became public domain during the legal investigation into the Enron corporation. The raw corpus contains approximately 600,000 emails; however we use the cleaned version provided by Fiore and Heer (2004), which consists of approximately 200,000 emails (70 million words). We remove HTML emails, and those with artefacts of conversions between encoding systems.

3. Annotations

3.1. Number Senses

Our previous work (Moore et al., 2009) on semi-supervised Number Sense Disambiguation used the 12 senses defined by Sproat et al. (2001) (see table 1 for a description of these) on the subsection of the North American News Text Corpus they had annotated. Our error analysis found some senses were too coarse grained, and some of the corpus labelling was inconsistent. 57% of numbers in the corpus were labelled with "NUM" (cardinal number) when we felt separating some of these senses would be beneficial.

We therefore present a more comprehensive set of 26 number senses, suitable for any domain, listed in table 2, which

Label		Description	Examples
NUM	(57%)	Number (Cardinal)	12, 45, 1/2, 0.6
NYER	(20%)	Year(s)	1998, 80s, 1900s
NORD	(8.7%)	Number (Ordinal)	May 7, 3rd, Bill Gates III
MONEY	(7.7%)	Money (US or other)	$3.45, HK$300, Y20,000, $200K
NIDE	(2.7%)	Identifier	747, 386, I5, pc110
NTEL	(1.4%)	Telephone number	212 555–4523
NTIME	(1.2%)	a (compound) time	3:20, 11:45
NDATE	(0.82%)	a (compound) date	2/2/99, 14/03/87 (or US) 03/14/87
NDIG	(0.20%)	Number as digits	Room 101
NADDR	(0.18%)	Building Number	5000 Pennsylvania Ave, 28 Kings Parade
NZIP	(0.18%)	Zip code or PO box	91020
PRCT	(0.06%)	Percentage	75%, 3.4%

Table 1: Number Senses defined in Sproat et al. (2001) (Percentages indicate frequency in the subset of the North American News Text Corpus)

we used within our annotation work. We sought to correct some of the difficulties we previously found with the Sproat et al. definitions - e.g. *'March 11'* being labelled an Ordinal and not as a Day of Month. The list of senses was derived by performing a series of trial annotations on both the Enron corpus and Sproat's newswire corpus, and adjusting senses accordingly. There is a mapping back from our new senses to those of Sproat et al.

Numbers within a URL, email address or attachment filename occur frequently within the email domain, but are likely to occur rarely in other contexts. In some cases, the numbers within a URL have an associated sense (*'www.london2012.com'*) - but in others, there is no obvious sense associated with the number. This was tackled using a two-step process - first the entire token (delimited by whitespace) is labelled with the sense URL (or email or filename); then individual numbers (contiguous digit strings) within the token are labelled with an appropriate sense.

3.2. Context

We hypothesise that human annotators can provide useful context hints for sense disambiguation systems, by being asked to highlight the neighbouring words that help them make their decision about each number's sense. For example, in the sentence fragment *'the estimated cleanup cost remaining is $6 million'* the dollar sign indicates that the number is a currency, but the word *'cost'* also suggests it. We plan to test this hypothesis in the future, and therefore have asked annotators to highlight the context surrounding numbers, and measured the time taken per annotation. This can be compared to the time taken to perform similar annotation without highlighting context, allowing investigation of whether n hours of annotation with context results in a more accurate system than n hours of annotation without context.

Sense	**Mapped Sense**	**Guidelines**
Day of Month	NORD	Includes 23rd November, April 6, 23-25 April, 1st of the month
Duration	NUM	60 minutes, 90 days, 30 years., 3 to 4 days, 45-50 minutes, 3-year.
Month	NUM	Anywhere a month is written as a number on its own (not in a short date)
Time	NTIME	Includes 3 o'clock, 5 past 12 (classify both numbers as time), 5:09, 13:00-14:00
Year	NYEAR	1998, '98, 98, 1976-7, 1980s
Quarter Year	NUM	Q1, Q3, Q4-08 Q2-2009
Timezone offset	NUM	-0700 or +0400 or +4hrs
Short Date	NDATE	23.4.09 16/4, Aug-2008, and also American equivalents. Annotators then select which of the following orderings best matches the short date: Day Month, Day Month Year, Month Day, Month Day Year, Year Month, Year Month Day, Year Day Month, or Unclear
Ordinal	NORD	1st, 3rd, 84th. Does not include Day of Month
Cardinal	NUM	3 people, 17 companies, 9 out of 10 cats, 5.4% year, 30mph, 16 kg, 800Mb. Includes numbers within sums (but not fractions). Does not include currency.
Chemical Formula	NUM	H2SO4, 2NaCl etc. Only use for molecular formulae. If used within a name, e.g. "2-Bromo-1-chloropropane" that should be labelled as an ID.
Fraction	NUM	1/2 1/4 56/120 Does not include whole numbers that are followed by fractions, e.g. the 1 in '1 3/4'
Currency	MONEY	30, $50, 60 GBP, 40 dollars, 10p, 30
Sport Score	NUM	A score from any sport.
Numbered List	NUM	1) 2) 3) (or 1. 2. 3. or similar)
ID	NID	An ID number or reference, referring to a particular object, class of objects, logical grouping etc. Includes order numbers, account numbers, car registrations, Room numbers, passport numbers. Includes software version numbers (e.g. Firefox 2.0) except for those that are years (Windows 98, Word 2000).
Idiom	NUM	Number with a special meaning, that would not persist if the digits were changed, e.g. 24/7. Does not include '999', '911' these are phone numbers. Does not include significant dates e.g. 9/11.
House Number	NADDR	House numbers, flat numbers etc. e.g. 54 Kings Parade, Flat 5 Does not include postal/zip codes, or examples such as '5th Avenue' - that's an Ordinal.
Post Code/ZIP	NZIP	CB2, 90210. Use for equivalent codes in any country.
Tel. Number	NTEL	Includes dialling codes, country codes, extensions. Includes fax and mobile/cell phones, and pagers, including "Pager Number xxx" Does not include 'press 1 for option a' etc.
Map Position	NUM	Latitude and Longitude, OS Grid reference, or any other map co-ordinate system. Include references only giving one co-ordinate, e.g. 'The Greenwich Meridian is at longitude 0'.
Numbered Road	NUM	Roads with associated numbers, e.g. M25, A1, Interstate 30, Route 66. Does not include 5th Avenue (that's an ordinal)
Email	NIDE	Email address
URL	NIDE	Part of a URL
Filename	NIDE	Part of a filename
Unclear	NUM	For numbers that don't clearly fit into any of the above categories

Table 2: Our Number Senses, their annotation guidelines, and their mapping to the senses used by Sproat et al.

3.3. Annotation Method

We developed a web interface, which annotators use to view emails, assign a sense to the target number, and highlight useful context words. The interface times how long was spent on each annotation, and users have the ability to pause the timer while they are performing other tasks. Annotations proceed one email at a time, so once annotators understand the topic of an email they can be quicker for later annotations.

To provide a balance between corpus coverage and allowing assessment of inter-annotator agreement, annotators are allocated emails such that:

- 20% of emails are annotated by all annotators (with all annotators highlighting context)
- 40% are annotated by just one pair of annotators (allowing pair wise comparison of inter-annotator agreement). For each of these emails, one annotator will be able to highlight context, and the other will not, allowing an analysis of the additional time needed to highlight the context.
- 40% are annotated by just one annotator, who will be able to highlight context for half of these emails, but not the other half - again allowing analysis of the time required to highlight the context.

Within the corpus, all numbers (including "words" containing a digit) are annotated within the subject or body of an email. The one exception is numbers within the most common email reply header (a block starting '——Original message——'). They are very frequent, have a regular, predictable format and would not lead to useful annotated data.

The decision to annotate one entire email at a time, rather than just some of the numbers within it, was taken to allow the user to label more quickly once they have already established the context. Frequently numbers occur next to each other (*'9 April 2009'*) and so the annotator had already established the sense of the next number as well. This was occasionally frustrating to annotators - in particular, an email advertising a website contained the string *'MP3.com'* 16 times, and each had to be annotated individually.

4. Corpus Statistics

We continue to annotate the corpus, but at present 28 annotators have annotated 1109 different numbers (some are annotated by more than one person; 2024 annotations have been made in total). Annotations have a pairwise inter-annotator agreement of 0.86, and a Krippendorff-α (Krippendorff, 1980) measure of 0.674.

The sense frequencies of the most common senses are shown in table 3. The distribution is very different to that of the Sproat et al. (2001) newswire text - in particular, times and telephone numbers are far more frequent in emails than in newswire text. To fairly evaluate the performance of NLP systems that use numbers on business email (a key 'real world' domain), these differences need to be taken into account.

4.1. Time Required to Annotate Context

Annotating context is only helpful if the extra time spent on annotation leads to a performance improvement, compared to the extra data that could have been labelled without context in this time.

We measured the time taken to annotate each sentence; the distribution is shown in figure 2. There were some extreme outlying durations in the data (in some cases, annotations that apparently took several hours); since annotations were web based, and volunteers were annotating in their free time, these are probably due to the annotator doing another task and coming back later. These distort the mean

Figure 1: The Web Based Annotation software. *'9,'* is being annotated. The annotator has highlighted *'April'* as providing the necessary context to establish the sense.

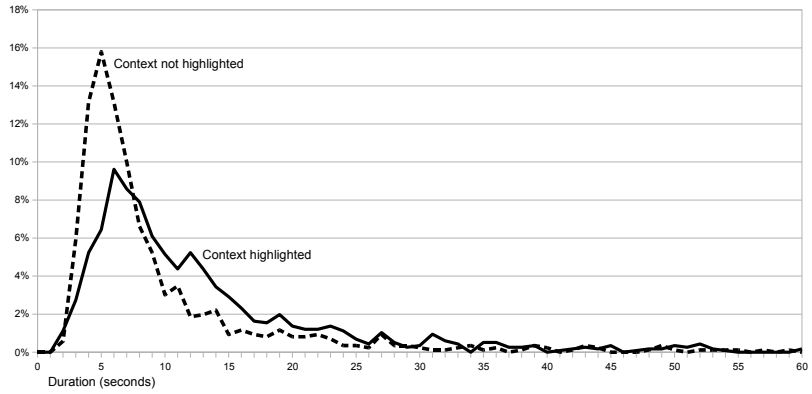

Figure 2: Distribution of annotation durations

Sense	frequency	NANTC
Time	11.1%	1.2%
Currency	10.8%	7.7%
Tel. Number	10.2%	1.2%
Day of Month	9.4%	8.7%*
Year	8.9%	20.0%
Cardinal	8.4%	57.0%
Numbered List	6.1%	-
ID	5.2%	2.7%
Short Date - Month Day Year	5.1%	0.8%**

* Includes both days of the month and other ordinal senses.
** All forms of Short Date.

Table 3: Comparing the distribution of the 9 most frequent senses in our corpus with that of the nearest equivalent sense of the North American News Text Corpus (NANTC) annotated by Sproat et al. (2001)

Measure	Ordinary	With Context	Ratio
Mean	24.5 sec	40.6 sec	1.66
Median	6.11 sec	9.42 sec	1.54
90th percentile	21.61 sec	29.27 sec	1.35
Mean within 90th percentile	9.88 sec	13.60 sec	1.38
Mean of durations < 10 min	11.11 sec	15.13 sec	1.36

Table 4: Analysis of how much longer annotations with additional context annotation take compared to those without. *'Mean within 90th percentile'* signifies taking the mean of the lowest 90% of durations.

durations, so we also use other measures to compare the durations (see table 4). These measures suggest that annotating the surrounding context results in annotation taking approximately 1.4 times as long.

5. Conclusion

This paper has presented a domain corpus (focussed on business emails) annotated for number senses. After the annotation is complete, we will make this corpus publicly available. We have show that the distribution of numbers within this domain is very different to that within Newswire text. We have also suggested that non-expert annotators can provide useful data by indicating the context that informed their choice of sense, and provided data to allow this to be investigated in the future. Together with the previously released newswire corpus of Sproat et al. (2001), our corpus can provide useful material for the training, testing and domain adaptation of number sense disambiguation systems.

Acknowledgements

Stuart Moore was supported by the EPSRC CASE Studentship co-sponsored by Toshiba Research Europe. We would like to thank Prof. Alan Black of Carnegie Mellon University for providing the corpora labelled by Sproat et al. (2001). We would also like to thank the annotators for donating their time.

6. References

Andrew Fiore and Jeff Heer. 2004. UC Berkely Enron Email Analysis. `http://bailando.sims.berkeley.edu/enron_email.html`.

David Graff. 1995. North American News Text Corpus. Linguistic Data Consortium.

Bryan Klimt and Yiming Yang. 2004. Introducing the Enron Corpus. In *CEAS 2004 - First Conference on Email and Anti-Spam*, Mountain View, California, USA, July 30-31.

Klaus Krippendorff, 1980. *Content Analysis: An Introduction to Its Methodology*, chapter 12. Sage, Beverly Hills, CA.

Stuart Moore, Anna Korhonen, and Sabine Buchholz. 2009. Number Sense Disambiguation. In *Proceedings of the Conference of Pacific Association for Computational Linguistics (PACLING '09)*.

Richard Sproat, Alan W. Black, Stanley F. Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. 2001. Normalization of non-standard words. *Computer Speech & Language*, 15(3):287–333.

David Yarowsky. 1996. Homograph disambiguation in text-to-speech synthesis. In *Progress in Speech Synthesis*, pages 159–175. Springer-Verlag, New York.