Capturing Quantities in Text

Gina Cook
Engineering and Computer Science
Concordia University
Montréal, QC Canada
gina.c.cook@gmail.ca

Abstract

Contents

1 Introduction

2 Textual Entailment

2.1 World Knowledge and Old Information

A great deal of context is not stated explicitly in the texts that we read. Texts are usually considered to add to our existing knowledge, connect previously knowledge. While newspaper texts generally fall neatly into background knowledge contexts (criminal acts and sentencing, celebrity gossip, natural disasters, global politics) and often seem to provide most of the relevant information necessary to understand the context genres which depart toward philosophy (blogs) or humour (witty magazine articles) often leave the obvious required background knowledge unstated for comical effect.

Iftene 2008 developed a module in his Text Entailment system which took Named Entities from the text and searched Wikipedia to construct background knowledge about those named entities to be used in the task. Of course, it is expected that the hypotheses offered in Textual Entailment tasks are expected to be consistent with reality, and thus consistent with Wikipedia over use of background knowledge can claim that true Hypotheses such as Bangladesh is east of India supposedly follow from a text that discusses the exports of Bangladesh and discusses nothing of its geographical position. Thus, misclassifying the hypothesis as entailed by the text incorrectly allows the text to be offered in the justification of a question answering system which provides the users the information which it used to arrive at its answer.

2.2 Expectations and "Normality"

In addition to background knowledge a sense of what is "normal" or what follows expectedly from a previous event is used by humans when building inferences from text. For example, the hypothesis that *11 troops were deployed* in Example 1 below violates the sense of normality and would cause a human to double-read the text to be sure this was indeed the fact.

(1) **Eleven troops is an unexpectedly small number of troops, and therefore unlikely**

Britain deployed troops to Afghanistan shortly after the attacks of 11 September, 2001. Few then thought that British forces would still be in Afghanistan in far larger numbers seven years on, nor that they would be involved in some of the fiercest fighting British forces have seen in decades, as part of Nato's International Security and Assistance Force (ISAF).

Britain has **11 troops** that take part in Nato's International Security and Assistance Force.

Resources such as VerbOcean 2004 identifies relations between verbs such as *enablement* and *happens-before* as shown in Table 2 below.

(2) VerbOcean provides relations between verbs such as Enablement and Happens-Before

Semantic Relation	Example	Transitive	Symmetric	Number in VerbOcean
Similarity	produce::create	Y	Y	11,515
Strength	wound :: kill	Y	N	4,220
Antonyms	open :: close	N	Y	1,973
Enablement	fight :: win	N	N	393
happens-before	buy :: own; marry :: divorce	Y	N	4,205

VerbOcean 2004, cited in Iftene 2008:46.

2.3 New Information and Discourse Representation

Generally considered the main source of information, the text itself presents new information which can be represented in the Discourse Representation framework of linguists Kamp and Reyle 1996.

2.3.1 Entities and Named Entities

Noun phrases differ in their specificity and their definiteness. The secificity and definiteness of noun phrases can usually be deduced from the articles, quantifiers or plurality of the nouns. In order to add information about entities and named entities in the discourse representation of a text it is important to resolve pronouns according to gender as shown in Example 3 below.

(3) Co-reference resolution with gendered discourse referents: **She** was finally found at 10.35am the next day by PC David Lloyd George. Something had caught his eye in the undergrowth - **her** blue school coat or the deep red of the gloves that **she** had been wearing - and he stepped off the forestry track. **Muriel Drinkwater** was lying on **her** back, one arm outstretched by **her** side, the other slightly raised but with the gloved hand limp. **Her** eyes were open, but there was no doubt that **she** was dead. **She** was just a few weeks short of **her** 13th birthday. The constable blew sharply on his whistle. The year was 1946, when murder was supposedly less commonplace than now. The killing of **Muriel Drinkwater** was to make headlines for months and would vex Detective Chief Inspector William "Bulldog" Chapman, the Scotland Yard detective who led the inquiry, until his death nine years later.

Muriel Drinkwater was killed in 1946.

Creating co-reference chains is more complicated when there are multiple possible references, and worse still when the references are very similar as the two skyscrapers discussed in Example 4 below.

(4) Co-reference chain with highly similar discourse referents: Seoul City said Monday**a 690-meter-tall, 133-story multifunctional skyscraper** will be constructed in Sangam-dong. Once built, it will be **the second highest** after *the 800-meter-high Burj Dubai*, which is under construction, by South Korean developer Samsung C&T. **The construction** will cost more than 3.3 trillion won ($2.37 billion), the city estimates. To raise funds, 23 local developers signed an MOU at a Seoul hotel Monday with Seoul Mayor Oh Se-hoon attending. "**The landmark building** will help make Seoul more attractive and become a new tourist attraction here," Oh said. **The multifunctional building** will have hotels, offices, department stores, convention centers and various recreational facilities including an aquarium and movie theaters.

The highest skyscraper in the world is being built in Dubai.

2.3.2 Closure and Disjoint Sets, exaustive

(5) Los Angeles County probation officials say they are now studying how other counties recover juvenile detention costs, after admitting they mistakenly billed parents for days when youths were held in probation camps and halls. By law, California counties can bill parents and legal guardians for some daily costs of detaining youths, but **only those whose parents** can afford to pay. Last year, more than 20,000 youths were admitted to probation camps and halls, and L.A. County billed parents a daily charge of $11.94 for camps, 23.63 for halls.

In Los Angeles County **all parents** have to pay the detention costs of their children.

2.4 Certainty and Probability

We are rarely certain of what we conclude, when doing automatic inferences from text certainty must be dealt with. Iftene 2008 determines that infinitives which follow verbs/predicates such as *glad, believe, claim, necessary, compulsory, free, attempt, trial, refuse, deny, ignore, plan, intend, propose, be able* are often less certain than verbs/predicates in other contexts.

This observation by Iftene mirrors the *epistemic* syntactic feature in linguistics. It is generally accepted that embedded verbs such as infinitives do have the same truth conditional semantics as main verbs. The semantics of modals, epistemic verbs and other verbs which take embedded verbs is often captured in the framework of Intensional Semantics (von Fintel and Hiem 2002).

Ultimately the degree of certainty of a verb/predicate depends on syntactic factors such as whether it is embedded, whether it is negated, whether it is qualified with an adverb, and of course

whether it is typically used by speakers in an uncertain or certain fashion in that particular context (as can be seen in Example 6 below). It is difficult to capture the certainty of a verb simply by comparing the verb/predicates in the text and hypothesis without doing a deeper analysis and dealing with the logical form underneath the text. Fortunately in many genre's which favour objectivity clear factual active sentences are more frequent then sentences which express uncertainty or hint at uncertainty.

(6) Sentences are rarely simple facts

Simple verbs speed, speak, spend, reach, raise, held, fled, lost, knew, paid *Tata* **paid** *investors 608 pence a share*

Mental verbs tried, believed, expected, thought, planned, intended, proposed, enjoyed, glad *Sepulcher, traditionally* **believed to mark** *the site where Jesus was crucified, buried and then resurrected.*

Raising verbs seemed, looked like *Those who want to save Acre's forest* **seem to be winning**.

Control verbs promised, tried *But an O2 insider said there had been problems with a companies* **trying to sell** *the pass codes.*

Acc+Inf or ECM verbs asked, wanted, enabled, expected, encouraged, pushed, forced, made, let, allowed *The government now* **wants to use** *the 450 MHz band to create a new digital wireless network that* **would cover** *the entire country.*

Modal verbs might, can/could, will/would, ought to should *The voluntary recall is considered a Class II recall since it covers products that* **might cause** *a temporary health problem or* **pose** *only a slight threat of a serious nature, the FDA said.*

Epistemic verbs must, has to *In Los Angeles County all parents* **have to pay** *the detention costs of their children.*

Counterfactuals would, might, attempted, denied *A teenager slashed a woman to death as he* **attempted to steal** *$6 off her–to buy beer.*

Evidential verbs heard, saw, said, mentioned, reported *The nephew of Elio Amato, an Italian mafia boss, is* **reported to have undergone training** *at a paramilitary camp in Bulgaria.*

Aspectual verbs started, finished, kept on, continued to *Kaplinsky first* **started to work** *for the BBC in 2002 after a two year tenure with Sky News. The US government* **wants to keep** *drug prices* **down**.

3 Recognizing Textual Entailment Challenge

The RTE challenge began in 2005 and was in its fifth year in 2009. Sammons et al 2010 identify RTE and other similar 'grand' challenges as the key to identifying the components needed in more complex tasks which are ultimately more useful than simple search engines and improved word counting and indexing algorithms.

The RTE challenge has evolved to use longer texts which require co-reference resolution, identification of contradiction, implicit and explicit negation and more advanced inferencing techniques which mirror closer the real world needs in inferencing. Four tasks which serve to benefit from inferencing services are listed in § 3.1 below. The RTE challenge originally began with rather shallow processing recources which relied on bag of word models and became gradualy more complex as exemplified by the paraphrasing system in Iftene 2008 in § 3.2 and the discourse system in Hickl and Bensely in § 3.3.

3.1 Uses of Textual Entailment

As one of the most linguistically challenging tasks in natural language processing recognizing textual entailments is part of tasks such as question answering and information retrieval where information must be extracted and summarized.

3.1.1 Question Answering

Question answering is the task of retrieving information from a single or a variety of sources to satisfy a user's question. Question answering show in Example 7 below can be as simple as finding an appropriate sentence which answers the question, or as complex as compiling the information from multiple sentences in multiple documents. Question Answering essentially is one service which can be provided if information retrieval § 3.1.2, information extraction § 3.1.3, and summarization § 3.1.4 can be provided.

(7) How many employees does Kaiser have?

Kaiser's 156,000 employees will be eligible for a pilot program connecting the Oakland health maintenance organization's health records with Microsoft's HealthVault, a free, Web-based medical database the company began in October.

Kaiser has at least 156,000 workers.

3.1.2 Information Retrieval

Information Retrieval is a task where relevant information is sorted and retrieved for the user. The text in Example 8 might be retrieved for a query string of *detain days terror suspect*.

(8) Retrieve texts relevant to user queries

Gordon Brown today made a defiant stand in defence of his plan to detain terror suspects for up to 42 days without charge, insisting he had balanced the need to protect civil liberties against the interests of national security. Police are allowed to detain terror suspects for 42 days.

Police can detain terror suspects for 42 days without charge.

3.1.3 Information Extraction

Information Extraction is a task which can be used to process unstructured text to build a knowledge base. In Example 9 below the fact *Kingston is 90 miles north of New York City* is extracted from a more general text which discusses a shopping mall in Kingston, and happens to mention the city's location with respect to New York City.

(9) Extract facts/data from unstructured text

The Hudson Valley Mall in Kingston, New York, U.S. is under lockdown after at least one man entered the building and opened fire. The mall is located about 90 miles north of New York City.
Kingston is found 90 miles north of New York City.

3.1.4 Summarization

Summarization is a task which can take as input one document or multiple document and produce a summary of the documents contents. At its simplest summarization can be extracting key sentences from teh document. Summarization is often considered more useful if it uses natural langauge generation and actually summarizes the information in the document(s) as shown in Example 10 below.

(10) Summarization can provide a more concise and objective statement.

More than a hundred protestors gathered Saturday in San Francisco to object to the Human Rights Campaign's stance on gender identity. The protestors picketed the HRC's annual fundraising dinner to protest the organization's efforts to remove gender identity protections from legislation that would add sexual orientation to the Employment Non-Discrimination Act (ENDA).
Over 100 people protested outside a HRC dinner against the exclusion of gender identity protection from ENDA.

3.2 Paraphrasing Resources - Iftene 2008

Iftene used a variety of resources to develop paraphrasing functionality for his textual entailment system. In a Grid computing paradigm he used a

- Lemmatizing service which found the lemmas of the words,
- Synonym service which used WordNet, eXtended WordNet to find a set of synonyms for the verbs
- Antonym service which used WordNet and verb ocean to find a set of antonyms to allow for passive-like transformations over the verbs
- Dirt service which used Dirt to find a list of similar verbs for paraphrasing
- Acronym service which finds potential full forms for acronyms
- Background Knowledge service which starts with a Named Entity and searches Wikipedia for sentences that contain the Named Entity to build a network of Ideas about that entity.

Iftene found that the WordNet resource, negation rules, Named Entity rules, and Contradiction rules accounted for most of his system's success as 2nd position in the two-way task (72.1% accuracy) and first position (68.5% accuracy) in the three-way task of RTE4 challenge in 2008.

3.3 Discourse Commitments - Hickl & Bensley 2007

4 Quantity Annotator

4.1 Exact Quantity

4.2 Approximate Quantity

4.3 Downward Entailing Contexts

5 Simple Inferencer

6 Results

7 Discussion

References

Hickl, A. and Bensley, J. 2007. A Discourse Commitment-Based Framework for Recognising Textual Entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. Pages 185-190. 28-29 June, Prague, Czech Republic.

Kamp, H & U Reyle, 1996 'A Calculus for First Order Discourse Representation Structures,' *Journal of Logic, Language and Information*, 5, 297348.

von Fintel, K and I Heim. 2002. *Lecture Notes on Intensional Semantics*. Ms. Massachusetts Institute of Technology.

Zweig, E. 2009. Number-Neutral Bare plurals and the Multiplicity Implicature. *Linguistics and Philosophy* Volume 32, Number 4, 353-407.

A Gate Implementation - Quantity Grammar

B Gate Implementation - Quantity Inferencer