THE UNIVERSITY OF SYDNEY

Using Volunteer Contributions from Non-Linguists to Address the Knowledge Bottleneck in NLP Systems

James W. D. Constable *Supervisor: James R. Curran*
School of Information Technologies
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

IN A NUTSHELL

Natural Language Processing (NLP) Systems typically require large amounts of training data to build accurate models of human language. Unfortunately, obtaining such data is an expensive and time-consuming task, due to both the sheer amount of data involved, and the high level of linguistic expertise required.

My project looks at easing this process by leveraging a currently under-utilised resource: regular people.

Most people are not used to talking about their native language in the same technical way a linguist would, yet they have no problem speaking and understanding it, and can quite easily identify sentences that are ungrammatical or illogical. They are, for all intents and purposes, experts.

I plan to explore ways of manipulating these linguistic tasks into a form that non-linguists are able to understand, and then use crowdsourcing techniques to focus our new workforce on solving specific linguistic problems that currently require the expertise of trained linguistic annotators.

MOTIVATION

What are corpora? Why are they important in NLP?

- A corpus is just a structured collection of text, usually annotated in some way to make it suitable for various processing tasks.
- The annotation schemes used in corpora vary depending on their purpose, but commonly include information like part-of-speech (POS) tags (nouns, verbs, etc.) and phrase structure.
- Corpora provide NLP researchers with representative samples of a language, useful for training various machine learning algorithms. The complex nature of natural language means that these tasks require massive amount of data to overcome problems of ambiguity and data sparsity, so there is demand for large, quality corpora.

Why are corpora so expensive to produce?

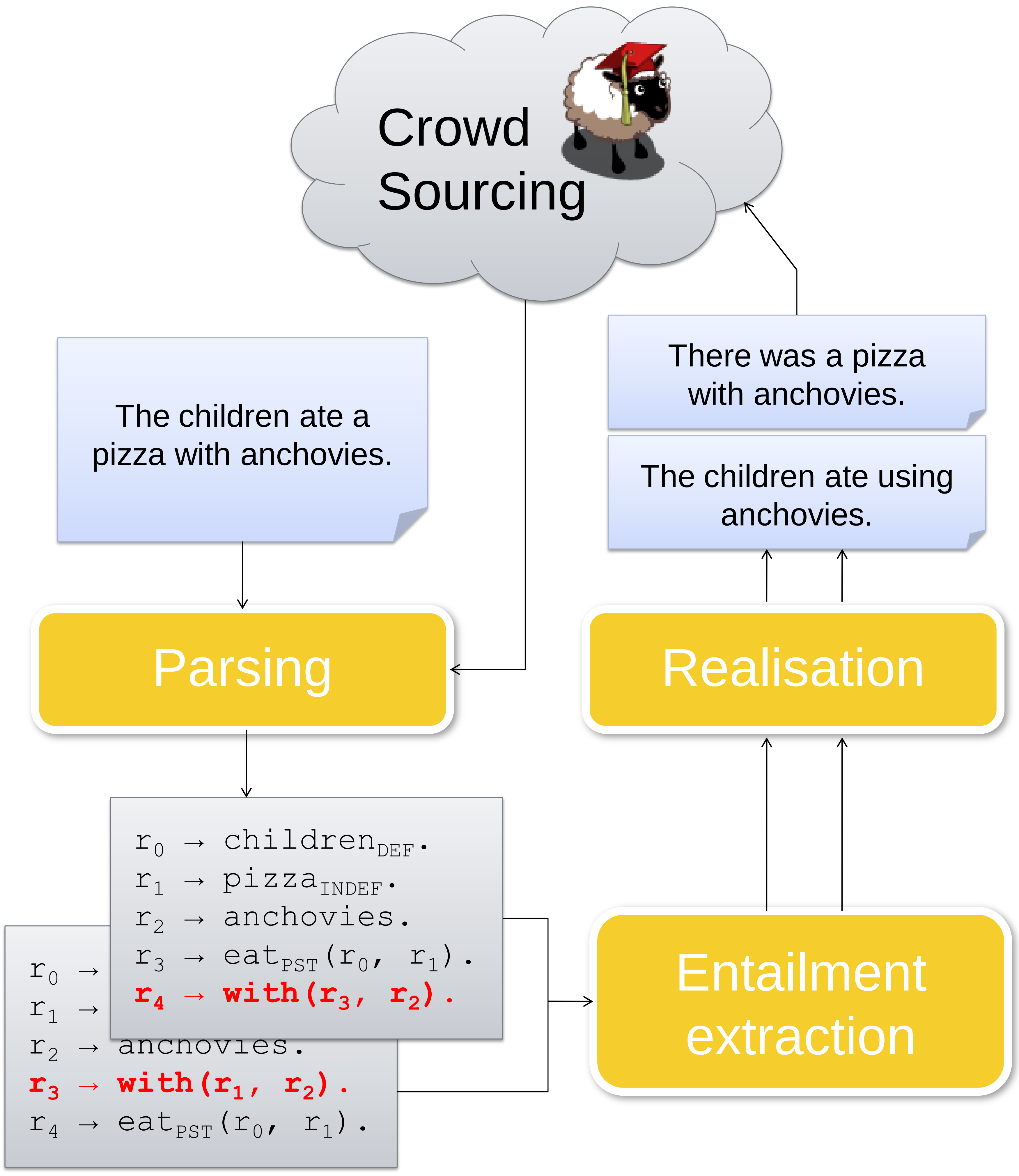
- Producing these annotations is no trivial task – it usually requires skilled annotators who are familiar with both the field and the annotation process, and can take many months depending on the size of the corpus. There is also the problem that many of the formalisms popular in NLP differ significantly from those used by linguists, so finding a large enough workforce can be challenging, and some amount of training is usually required.
- Corpora also go out of date. The Wall Street Journal section of the Penn Treebank is one of the most widely used corpora for training and evaluating NLP systems, yet the data was collected over 20 years ago (1987-1989), and consequently contains no URLs or email addresses. Considering the prevalence of these forms in modern text, we cannot really expect systems trained on such dated data to perform to their maximum potential.

THE PROJECT

- There are two main challenges in this project: the problem of reworking linguistic tasks to make them understandable to regular people, and the problem of finding the best medium for crowdsourcing such a project.
- To rework the tasks, we are currently looking at using a combination of parsing and realisation. Parsing can be seen as the problem of taking a piece of text and analysing it to determine its structure, so that the underlying meaning (logical form) can be uncovered. Realisation is the opposite problem: given a logical form, generating a human readable surface form. To date, there has been no real research into systems that integrate both, so in this respect our system will be one of the first of its kind. As the majority of our work is couched within the CCG formalism (Combinatory Categorial Grammar; Steedman (2000)), we will be using the C&C parser (Clark and Curran, 2007), and the OpenCCG realiser (White et al, 2007) as our parsing and realisation subsystems respectively.
- For the crowdsourcing component of this research, it is necessary to find a way to motivate users to help with our project. We see two primary ways in which this could be achieved – either with a monetary reward, or by appealing to users' sense of enjoyment or altruism. Both methods have previously been used academically, with mixed results (Kittur et al, 2008; von Ahn and Dabbish, 2008; Chamberlain et al., 2008). As a large part of our original motivation for this topic was to be able to gather data *cheaply*, we are currently focusing on the second approach, looking towards fashioning our task as an online game.

THE GRAND PLAN

The figure below gives a schematic summary of what we hope to achieve with this project. When the parser encounters a troublesome sentence (in this case, a PP-attachment ambiguity), it responds by passing its generated logical form(s) to the entailment extraction process, which manipulates them to target the problem at hand. In this case there is a conflict in the logical forms (highlighted in red), so the entailment extractor targets this in its output. The realiser then produces human-readable surface forms from the logical entailments, and presents these to the crowd sourcing system for human (dis)approval. The results are then fed back into the parsing system, so it can learn from its earlier mistakes.


Crowd Sourcing
There was a pizza with anchovies.
The children ate a pizza with anchovies.
The children ate using anchovies.
Parsing
Realisation
Entailment extraction

$r_0 \rightarrow \mathrm{children}_{DEF}.$
$r_1 \rightarrow \mathrm{pizza}_{INDEF}.$
$r_2 \rightarrow \mathrm{anchovies}.$
$r_3 \rightarrow \mathrm{eat}_{PST}(r_0, r_1).$
$\mathbf{r_4 \rightarrow with(r_3, r_2).}$

$r_0 \rightarrow$
$r_1 \rightarrow$
$r_2 \rightarrow$ anchovies.
$\mathbf{r_3 \rightarrow with(r_1, r_2).}$
$r_4 \rightarrow \mathrm{eat}_{PST}(r_0, r_1).$

REFERENCES

- Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. Phrase Detectives: A Web-based Collaborative Annotation Game. *Proceedings of I-Semantics, Gruz*, 2008.
- Stephen Clark and James R. Curran. Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Models, *Computational Linguistics*, 33(4):493-552, 2007.
- Kittur, A. and Chi, E.H. and Suh, B. Crowdsourcing user studies with Mechanical Turk. *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, 453-456, ACM, 2008.
- Mark Steedman. *The Syntactic Process*. Massachusetts Institute of Technology, USA, 2000.
- Luis von Ahn and Laura Dabbish. Designing Games With A Purpose. *Communications of the ACM*, 51(8):58-67, 2008.
- White, M. and Rajkumar, R. and Martin, S. Towards broad coverage surface realization with CCG. *Proceedings of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT)*, Citeseer, 2007.