Cognitively Plausible Models of Human Language Processing

Frank Keller
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
keller@inf.ed.ac.uk

Abstract

We pose the development of cognitively plausible models of human language processing as a challenge for computational linguistics. Existing models can only deal with isolated phenomena (e.g., garden paths) on small, specifically selected data sets. The challenge is to build models that integrate multiple aspects of human language processing at the syntactic, semantic, and discourse level. Like human language processing, these models should be incremental, predictive, broad coverage, and robust to noise. This challenge can only be met if standardized data sets and evaluation measures are developed.

1 Introduction

In many respects, human language processing is the ultimate goldstandard for computational linguistics. Humans understand and generate language with amazing speed and accuracy, they are able to deal with ambiguity and noise effortlessly and can adapt to new speakers, domains, and registers. Most surprisingly, they achieve this competency on the basis of limited training data (Hart and Risley, 1995), using learning algorithms that are largely unsupervised.

Given the impressive performance of humans as language processors, it seems natural to turn to psycholinguistics, the discipline that studies human language processing, as a source of information about the design of efficient language processing systems. Indeed, psycholinguists have uncovered an impressive array of relevant facts (reviewed in Section 2), but computational linguists are often not aware of this literature, and results about human language processing rarely inform the design, implementation, or evaluation of artificial language processing systems.

At the same time, research in psycholinguistics is often oblivious of work in computational linguistics (CL). To test their theories, psycholinguists construct computational models of human language processing, but these models often fall short of the engineering standards that are generally accepted in the CL community (e.g., broad coverage, robustness, efficiency): typical psycholinguistic models only deal with isolated phenomena and fail to scale to realistic data sets. A particular issue is evaluation, which is typically anecdotal, performed on a small set of handcrafted examples (see Sections 3).

In this paper, we propose a challenge that requires the combination of research efforts in computational linguistics and psycholinguistics: the development of cognitively plausible models of human language processing. This task can be decomposed into a modeling challenge (building models that instantiate known properties of human language processing) and a data and evaluation challenge (accounting for experimental findings and evaluating against standardized data sets), which we will discuss in turn.

2 Modeling Challenge

2.1 Key Properties

The first part of the challenge is to develop a model that instantiates key properties of human language processing, as established by psycholinguistic experimentation (see Table 1 for an overview and representative references).\footnote{Here an in the following, we will focus on sentence processing, which is often regarded as a central aspect of human language processing. A more comprehensive answer to our modeling challenge should also include phonological and morphological processing, semantic inference, discourse processing, and other non-syntactic aspects of language processing. Furthermore, established results regarding the interface between language processing and non-linguistic cognition (e.g., the sensorimotor system) should ultimately be accounted for in a fully comprehensive model.}

A striking property of the human language processor is its efficiency and robustness. For the vast majority of sentences, it will effortlessly and rapidly deliver the correct analysis, even in the face of noise and ungrammaticalities. There is considerable experimental evi-
dence that shallow processing strategies are used to achieve this. The processor also achieves broad coverage: it can deal with a wide variety of syntactic constructions, and is not restricted by the domain, register, or modality of the input.

Human language processing is also word-by-word incremental. There is strong evidence that a new word is integrated as soon as it is available into the representation of the sentence thus far. Readers and listeners experience differential processing difficulty during this integration process, depending on the properties of the new word and its relationship to the preceding context. There is evidence that the processor instantiates a strict form of incrementality by building only fully connected trees. Furthermore, the processor is able to make predictions about upcoming material on the basis of sentence prefixes. For instance, listeners can predict an upcoming post-verbal element based on the semantics of the preceding verb. Or they can make syntactic predictions, e.g., if they encounter the word either, they predict an upcoming or and the type of complement that follows it.

Another key property of human language processing is the fact that it operates with limited memory, and that structures in memory are subject to decay and interference. In particular, the processor is known to incur a distance-based memory cost: combining the head of a phrase with its syntactic dependents is more difficult the more dependents have to be integrated and the further away they are. This integration process is also subject to interference from similar items that have to be held in memory at the same time.

### 2.2 Current Models

The challenge is to develop a computational model that captures the key properties of human language processing outlined in the previous section. A number of relevant models have been developed, mostly based on probabilistic parsing techniques, but none of them instantiates all the key properties discussed above (Table 1 gives an overview of model properties).

The earliest approaches were ranking-based models (Rank), which make psycholinguistic predictions based on the ranking of the syntactic analyses produced by a probabilistic parser. Jurafsky (1996) assumes that processing difficulty is triggered if the correct analysis falls below a certain probability threshold (i.e., is pruned by the parser). Similarly, Crocker and Brants (2000) assume that processing difficulty ensures if the highest-ranked analysis changes from one word to the next. Both approaches have been shown to successfully model garden path effects. Being based on probabilistic parsing techniques, ranking-based models generally achieve a broad coverage, but their efficiency and robustness has not been evaluated. Also, they are not designed to capture syntactic prediction or memory effects (other than search with a narrow beam in Brants and Crocker 2000).

The ranking-based approach has been generalized by surprisal models (Surp), which predict processing difficulty based on the change in the probability distribution over possible analyses from one word to the next (Hale, 2001; Levy, 2008; Demberg and Keller, 2008a; Ferrara Boston et al., 2008; Roark et al., 2009). These models have been successful in accounting for a range of experimental data, and they achieve broad coverage. They also instantiate a limited form of prediction, viz., they build up expectations about the next word in the input. On the other hand, the efficiency and robustness of these models has largely not been evaluated, and memory costs are not modeled (again except for restrictions in beam size).

The prediction model (Pred) explicitly predicts syntactic structure for upcoming words (Demberg and Keller, 2008b, 2009), thus accounting for experimental results on predictive language processing. It also implements a strict form of incre-

---

<table>
<thead>
<tr>
<th>Property</th>
<th>Evidence</th>
<th>Model Rank</th>
<th>Model Surp</th>
<th>Model Pred</th>
<th>Model Stack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Efficiency and robustness</td>
<td>Ferreira et al. (2001); Sanford and Sturt (2002)</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>Broad coverage</td>
<td>Crocker and Brants (2000)</td>
<td>+</td>
<td>+</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>Incrementality and connectedness</td>
<td>Tanenhaus et al. (1995); Sturt and Lombardo (2005)</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Prediction</td>
<td>Kamide et al. (2003); Staub and Clifton (2006)</td>
<td>–</td>
<td>±</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>Memory cost</td>
<td>Gibson (1998); Vasishth and Lewis (2006)</td>
<td>–</td>
<td>–</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Table 1: Key properties of human language processing and their instantiation in various models of sentence processing (see Section 2 for details).
mentality by building fully connected trees. Memory costs are modeled directly as a distance-based penalty that is incurred when a prediction has to be verified later in the sentence. However, the current implementation of the prediction model is neither robust and efficient nor offers broad coverage.

Recently, a stack-based model (Stack) has been proposed that imposes explicit, cognitively motivated memory constraints on the parser, in effect limiting the stack size available to the parser (Schuler et al., 2010). This delivers robustness, efficiency, and broad coverage, but does not model syntactic prediction. Unlike the other models discussed here, no psycholinguistic evaluation has been conducted on the stack-based model, so its cognitive plausibility is preliminary.

### 2.4 Acquisition and Crosslinguistics

All models of human language processing discussed so far rely on supervised training data. This raises another aspect of the modeling challenge: the human language processor is the product of an acquisition process that is largely unsupervised and has access to only limited training data: children aged 12–36 months are exposed to between 10 and 35 million words of input (Hart and Risley, 1995). The challenge therefore is to develop a model of language acquisition that works with such small training sets, while also giving rise to a language processor that meets the key criteria in Table 1. The CL community is in a good position to rise to this challenge, given the significant progress in unsupervised parsing in recent years (starting from Klein and Manning 2002). However, none of the existing unsupervised models has been evaluated against psycholinguistic data sets, and they are not designed to meet even basic psycholinguistic criteria such as incrementality.

A related modeling challenge is the development of processing models for languages other than English. There is a growing body of experimental research investigating human language processing in other languages, but virtually all existing psycholinguistic models only work for English (the only exceptions we are aware of are Dubey et al.’s (2008) and Ferrara Boston et al.’s
Again, the CL community has made significant progress in crosslinguistic parsing, especially using dependency grammar (Hajič, 2009), and psycholinguistic modeling could benefit from this in order to meet the challenge of developing crosslinguistically valid models of human language processing.

3 Data and Evaluation Challenge

3.1 Test Sets

The second key challenge that needs to be addressed in order to develop cognitively plausible models of human language processing concerns test data and model evaluation. Here, the state of the art in psycholinguistic modeling lags significantly behind standards in the CL community. Most of the models discussed in Section 2 have not been evaluated rigorously. The authors typically describe their performance on a small set of hand-picked examples; no attempts are made to test on a range of items from the experimental literature and determine model fit directly against behavioral measures (e.g., reading times). This makes it very hard to obtain a realistic estimate of how well the models achieve their aim of capturing human language processing behavior.

We therefore suggest the development of standard test sets for psycholinguistic modeling, similar to what is commonplace for tasks in computational linguistics: parsers are evaluated against the Penn Treebank, word sense disambiguation systems against the SemEval data sets, co-reference systems against the Tipster or ACE corpora, etc. Two types of test data are required for psycholinguistic modeling. The first type of test data consists of a collection of representative experimental results. This collection should contain the actual experimental materials (sentences or discourse fragments) used in the experiments, together with the behavioral measurements obtained (reading times, eye-movement records, rating judgments, etc.). The experiments included in this test set would be chosen to cover a wide range of experimental phenomena, e.g., garden paths, syntactic complexity, memory effects, semantic and discourse factors. Such a test set will enable the standardized evaluation of psycholinguistic models by comparing the model predictions (rankings, surprisal values, memory costs, etc.) against behavioral measures on a large set of items. This way both the coverage of a model (how many phenomena can it account for) and its accuracy (how well does it fit the behavioral data) can be assessed.

Experimental test sets should be complemented by test sets based on corpus data. In order to assess the efficiency, robustness, and broad coverage of a model, a corpus of unrestricted, naturally occurring text is required. The use of contextualized language data makes it possible to assess not only syntactic models, but also models that capture discourse effects. These corpora need to be annotated with behavioral measures, e.g., eye-tracking or reading time data. Some relevant corpora have already been constructed, see the overview in Table 3, and various authors have used them for model evaluation (Demberg and Keller, 2008a; Pynte et al., 2008; Frank, 2009; Ferrara Boston et al., 2008; Patil et al., 2009; Roark et al., 2009; Mitchell et al., 2010).

However, the usefulness of the psycholinguistic corpora in Table 3 is restricted by the absence of gold-standard linguistic annotation (though the French part of the Dundee corpus, which is syntactically annotated). This makes it difficult to test the accuracy of the linguistic structures computed by a model, and restricts evaluation to behavioral predictions. The challenge is therefore to collect a standardized test set of naturally occurring text or speech enriched not only with behavioral variables, but also with syntactic and semantic annotation. Such a data set could for example be constructed by eye-tracking section 23 of the Penn Treebank (which is also part of Propbank, and thus has both syntactic and thematic role annotation).

In computational linguistics, the development of new data sets is often stimulated by competitions in which systems are compared on a standardized task, using a data set specifically designed for the competition. Examples include the CoNLL shared task, SemEval, or TREC in computational syntax, semantics, and discourse, respectively. A similar competition could be developed for computational psycholinguistics – maybe along the lines of the model comparison challenges that held at the International Conference on Cognitive Modeling. These challenges provide standardized task descriptions and data sets; participants can enter their cognitive models, which were then compared using a pre-defined evaluation metric.\footnote{The ICCM 2009 challenge was the Dynamic Stock and Flows Task, for more information see \url{http://www.hss.cmu.edu/departments/sds/ddmlab/modeldsf/}.}

\[\]


<table>
<thead>
<tr>
<th>Corpus</th>
<th>Language</th>
<th>Words</th>
<th>Participants</th>
<th>Method</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dundee Corpus</td>
<td>English, French</td>
<td>50,000</td>
<td>10</td>
<td>Eye-tracking</td>
<td>Kennedy and Pynte (2005)</td>
</tr>
<tr>
<td>Potsdam Corpus</td>
<td>German</td>
<td>1,138</td>
<td>222</td>
<td>Eye-tracking</td>
<td>Kliegl et al. (2006)</td>
</tr>
</tbody>
</table>

Table 3: Test corpora that have been used for psycholinguistic modeling of sentence processing; note that the Potsdam Corpus consists of isolated sentences, rather than of continuous text.

3.2 Behavioral and Neural Data

As outlined in the previous section, a number of authors have evaluated psycholinguistic models against eye-tracking or reading time corpora. Part of the data and evaluation challenge is to extend this evaluation to neural data as provided by event-related potential (ERP) or brain imaging studies (e.g., using functional magnetic resonance imaging, fMRI). Neural data sets are considerably more complex than behavioral ones, and modeling them is an important new task that the community is only beginning to address. Some recent work has evaluated models of word semantics against ERP (Murphy et al., 2009) or fMRI data (Mitchell et al., 2008). This is a very promising direction, and the challenge is to extend this approach to the sentence and discourse level (see Bachrach 2008). Again, it will again be necessary to develop standardized test sets of both experimental data and corpus data.

3.3 Evaluation Measures

We also anticipate that the availability of new test data sets will facilitate the development of new evaluation measures that specifically test the validity of psycholinguistic models. Established CL evaluation measures such as Parseval are of limited use, as they can only test the linguistic, but not the behavioral or neural predictions of a model.

So far, many authors have relied on qualitative evaluation: if a model predicts a difference in (for instance) reading time between two types of sentences where such a difference was also found experimentally, then that counts as a successful test. In most cases, no quantitative evaluation is performed, as this would require modeling the reading times for individual item and individual participants. Suitable procedures for performing such tests do not currently exist; linear mixed effects models (Baayen et al., 2008) provide a way of dealing with item and participant variation, but crucially do not enable direct comparisons between models in terms of goodness of fit.

Furthermore, issues arise from the fact that we often want to compare model fit for multiple experiments (ideally without reparametrizing the models), and that various mutually dependent measures are used for evaluation, e.g., processing effort at the sentence, word, and character level. An important open challenge is there to develop evaluation measures and associated statistical procedures that can deal with these problems.

4 Conclusions

In this paper, we discussed the modeling and data/evaluation challenges involved in developing cognitively plausible models of human language processing. Developing computational models is of scientific importance in so far as models are implemented theories: models of language processing allow us to test scientific hypotheses about the cognitive processes that underpin language processing. This type of precise, formalized hypothesis testing is only possible if standardized data sets and uniform evaluation procedures are available, as outlined in the present paper. Ultimately, this approach enables qualitative and quantitative comparisons between theories, and thus enhances our understanding of a key aspect of human cognition, language processing.

There is also an applied side to the proposed challenge. Once computational models of human language processing are available, they can be used to predict the difficulty that humans experience when processing text or speech. This is useful for a number of applications: for instance, natural language generation would benefit from being able to assess whether machine-generated text or speech is easy to process. For text simplification (e.g., for children or impaired readers), such a model is even more essential. It could also be used to assess the readability of text, which is of interest in educational applications (e.g., essay scoring). In machine translation, evaluating the fluency of system output is crucial, and a model that predicts processing difficulty could be used for this, or to guide the choice between alternative translations, and maybe even to inform human post-editing.

---

4 These data sets were released as part of the NAACL-2010 Workshop on Computational Neurolinguistics.
References


