Available online at www.sciencedirect.com
ScienceDirect
Cognitive Systems Research 9 (2008) 198–213
Cognitive Systems RESEARCH
www.elsevier.com/locate/cogsys

Automatic evaluation of syntactic learners in typologically-different languages

Action editor: Gregg Oden

Franklin Chang [a,*], Elena Lieven [b], Michael Tomasello [b]

[a] *Cognitive Language Information Processing Open Laboratory, NTT Communication Sciences Laboratories, NTT Corp., 2-4 Hikari-dai, Seika-cho, Souraku-gun, 6190237 Kyoto, Japan*

[b] *Department of Developmental and Comparative Psychology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany*

Received 7 June 2007; received in revised form 12 September 2007; accepted 6 October 2007
Available online 1 November 2007

Abstract

Human syntax acquisition involves a system that can learn constraints on possible word sequences in typologically-different human languages. Evaluation of computational syntax acquisition systems typically involves theory-specific or language-specific assumptions that make it difficult to compare results in multiple languages. To address this problem, a bag-of-words incremental generation (BIG) task with an automatic sentence prediction accuracy (SPA) evaluation measure was developed. The BIG–SPA task was used to test several learners that incorporated *n*-gram statistics which are commonly found in statistical approaches to syntax acquisition. In addition, a novel Adjacency–Prominence learner, that was based on psycholinguistic work in sentence production and syntax acquisition, was also tested and it was found that this learner yielded the best results in this task on these languages. In general, the BIG–SPA task is argued to be a useful platform for comparing explicit theories of syntax acquisition in multiple languages.

Keywords: Syntax acquisition; Computational linguistics; Corpora; Syntax evaluation; Linguistic typology

1. Introduction

Children, computers, and linguists have similar challenges in extracting syntactic constraints from language input. Any system that acquires syntactic knowledge (a syntactic learner) must confront the fact that words do not come labeled with syntactic categories and the syntactic relations that can hold among these words can vary to a great extent among languages. This article presents a method for evaluating syntactic learners, that is, how well they have acquired syntactic knowledge from the input. This method, which uses a *bag-of-words incremental generation* (BIG) task and an evaluation measure called *sentence prediction accuracy* (SPA), is applied to several formally-specified learners, as well as to a new learner called the Adjacency–Prominence learner. It will be shown that the SPA measure is capable of evaluating the syntactic abilities in a variety of learners using input from typologically-different languages and it does so in a manner that is relatively free of assumptions about the form of linguistic knowledge.

Words in utterances are not labeled with syntactic categories, and there is variability in how linguistic theories characterize the syntactic constraints on an utterance. For example, constructions are a type of syntactic unit in some theories (Goldberg, 1995), but not others (Chomsky, 1981). Syntactic constraints also differ across languages, and it is difficult to adapt a particular theory of syntactic categories or constraints to typologically-different languages (Croft, 2001). For example, the adjective category is often thought to be a universal syntactic category, but in many languages, it is difficult to distinguish adjectives and stative verbs (e.g., Chinese, Li & Thompson, 1990)

* Corresponding author. Tel.: +81 774 93 5273; fax: +81 774 93 5345.
E-mail address: chang.franklin@gmail.com (F. Chang).

1389-0417/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.cogsys.2007.10.002

and in some languages, there are several adjective categories (e.g., Japanese, Tsujimura, 1996). Since the labeling of corpora requires that one make particular assumptions about the nature of syntax, the evaluation of syntactic knowledge with these human-labeled corpora is both theory- and language-dependent. These evaluation methods work best for mature areas of syntactic theory, such as the evaluation of adult English syntactic knowledge, but are less suited for areas such as syntax acquisition or linguistic typology, where there is more controversy about the nature of syntax (Croft, 2001; Pinker, 1989; Tomasello, 2003).

A large number of computational approaches for learning syntactic knowledge are evaluated against human-labeled corpora. For example in part-of-speech tagging, a tagger attempts to predict the syntactic category (or tag) for each of the words in an utterance, and the system is evaluated by comparing its output against the human-labeled tag sequence associated with the test utterance (Church, 1989; Dermatas & Kokkinakis, 1995). The set of tag categories that are used to label a particular corpus is called its tagset, and different corpora, even in the same language, use different tagsets (Jurafsky & Martin, 2000). In addition, the same tagger can show different levels of performance, when evaluated against different types of corpora or different tagsets. Atwell et al. (2000) trained a supervised tagger with a single corpus that had been tagged with eight different English tagsets and found significant variation among the tagsets in test accuracy from 86.4% to 94.3%. When taggers are applied to multiple languages, there is an additional problem that the tagsets are not equated across the languages, because tagsets can vary in the specificity of the categories or in the degree that semantic or formal criteria are used for assignment of categories (Croft, 2001). For example, Dermatas and Kokkinakis (1995) found that the same Hidden Markov Model for part-of-speech tagging (HMM-TS2) with the same amount of input (50,000 words) labeled with the same set of categories (extended grammatical classes) yielded better accuracy levels for English (around 5% prediction error, EEC-law text) than for five other European languages (Greek yielded more than 20% prediction error). Since many of the relevant factors were controlled here (e.g., input size, learner, categories), the large variability in accuracy is probably due to the match between the categories and the utterances in the corpora, in this case, the match was better for English than Greek. If that is the case, it suggests that evaluating these systems with this tagset is inherently biased towards English. Other evaluation measures in computational linguistics, such as the learning of dependency structures, also seem to be biased toward English. Klein and Manning (2004) found that their unsupervised dependency model with valence plus constituent-context learner yielded accuracy results in English of 77.6% (Fig. 6 in their paper, UF_1), but German was 13.7% lower and Chinese was 34.3% lower. In addition to these biases, English corpora are often larger and more consistently labeled and together these factors help to insure that there will be a bias towards English in evaluation of computational systems. But since humans can learn any human language equally well, it is desirable to have a way to evaluate syntax that is not inherently biased for particular languages.

One area of computational linguistics that has been forced to deal with variability in syntax across languages is the domain of machine translation. In translating an utterance from a source language to a target language, these systems attempt to satisfy two constraints. One constraint is to ensure that the meaning of the source utterance is preserved in the target utterance and the other constraint is that the order of words in the target utterance should respect the syntactic constraints of the target language. In statistical approaches to machine translation, these constraints are supported by two components: the translation model and the language model (Brown, Della Pietra, Della Pietra, & Mercer, 1993). The translation model assumes that the words in the source utterance capture some of its meaning, and this meaning can be transferred to the target utterance by translating the words in the source language into the target language. Since words in some languages do not have correspondences in other languages, the set of translated words can be augmented with additional words or words can be removed from the set. This set of translated words will be referred to as a bag-of-words, since the order of the words may not be appropriate for the target language. The ordering of the bag-of-words for the syntax of the target language is called decoding, and involves the statistics in the language model. Statistical machine translation systems are not able to match human generated translations, but they are able to generate translations of fairly long and complicated utterances and these utterances can be often understood by native speakers of the target language.

In statistical machine translation, the ordering of the words in an utterance is a whole utterance optimization process, where the goal is to optimize a particular metric (e.g., the transition probabilities between words) over the whole utterance. This optimization is computationally intensive, since finding an optimal path through a set of words is equivalent to the Traveling Salesman problem and therefore is NP-complete (Knight, 1999). There is however no guarantee that humans are doing whole sentence optimization of the sort that is used in statistical machine translation. And there is experimental evidence from humans that contradicts the assumptions of whole sentence optimization and suggests instead that speakers can plan utterances incrementally. Incremental planning means that speakers plan sentences word-by-word using various scopes of syntactic and message information. Incremental planning during production predicts that words that are more accessible due to lexical, semantic, or discourse factors will tend to come earlier in utterances and there is a large amount of experimental evidence supporting this (Bock, 1982, 1986; Bock & Irwin, 1980; Bock & Warren, 1985; Bock, Loebell, & Morey, 1992; Ferreira & Yoshita,