The Importance of Discourse Context for Statistical Natural Language Generation

Cassandre Creswell
Department of Linguistics
University of Toronto
creswell@cs.toronto.edu

Elsi Kaiser
Center for Language Sciences
University of Rochester
ekaiser@ling.rochester.edu

Abstract

Surface realization in statistical natural language generation is based on the idea that when there are many ways to say the same thing, the most frequent option based on corpus counts is the best. Based on data from English and Finnish, we argue instead that all options are not equivalent, and the most frequent one can be incoherent in some contexts. A statistical NLG system where word order choice is based only on frequency counts of forms cannot capture the contextually-appropriate use of word order. We describe an alternative method for word order selection and show how it outperforms a frequency-only approach.

1 Introduction

The purpose of a natural language generation (NLG) system is to encode semantic content in a linguistic form easily understood by humans in order to communicate it to the user of the system. Ideally, this content should be encoded in strings that are both grammatical and contextually appropriate.

Human speakers of all natural languages have many ways to encode the same truth-conditional meaning besides a single "canonical" word order, even when encoding one predicate and its arguments as a main clause. Humans choose contextually-appropriate options from these many ways with little conscious effort and with rather effective communicative results. Statistical approaches to natural language generation are based on the assumption that often many of these options will be equally good, e.g. (Bangalore and Rambow, 2000).

In this paper, we argue that, in fact, not all options are equivalent, based on linguistic data both from English, a language with relatively static word order, and from Finnish, a language with much more flexible word order. We show that a statistical NLG algorithm based only on counts of trees cannot capture the appropriate use of word order. We provide an alternative method which has been implemented elsewhere and show that it dramatically outperforms the statistical approach. Finally, we explain how the alternative method could be used to augment present statistical approaches and draw some lessons for future development of statistical NLG.

2 Statistical NLG: a brief summary

In recent years, a new approach to NLG has emerged, which hopes to build on the success of the use of probabilistic models in natural language understanding (Langkilde and Knight, 1998; Bangalore and Rambow, 2000; Ratnaparkhi, 2000). Building an NLG system is highly labor-intensive. For the system to be robust, large amounts of world and linguistic knowledge must be hand-coded. The goal of statistical approaches is to minimize hand-coding and instead rely upon information automatically extracted from linguistic corpora when selecting a linguistic realization of some conceptual representation.

The underlying concept of these statistical approaches is that the form generated to express a particular meaning should be selected on the basis of counts of that form (either strings or trees) in a corpus. In other words, in generating a form f to express an input, one wants to maximize the probability of the form, $P(f)$, with respect to some gold-standard corpus, and thus express the input in a way that resembles the realizations in the corpus most closely (Bangalore and Rambow, 2000). Bangalore and Rambow's algorithm for generating a string in the FERGUS system begins with an underspecified conceptual representation which is mapped to a dependency tree with unordered sibling nodes. To convert the dependency tree into a surface form, a syntactic structure is chosen for each node. In FERGUS, this structure is an elementary tree in a tree-adjoining grammar. The choice of a tree is stochastic, based on a tree model derived from 1,000,000

words of the Wall Street Journal. For example, the tree chosen for a verb V will be the most frequently found tree in the corpus headed by V.

3 Where counting forms fails

This section provides evidence from English and Finnish that word order affects meaning and acceptability. For each phenomenon we show how a statistical generation technique based only on the probability of forms in a corpus will fail to capture this distinction in meaning.

Speakers can use a particular form to indicate their assumptions about the status of entities, properties, and events in the discourse model. For example, references to entities may appear as full NPs, pronouns, or be missing entirely, depending on whether speakers regard them as new or old to the hearer or the discourse or as particularly salient (Gundel et al., 1993; Prince, 1992). Not just the lexical form of referential expressions, but also their position or role within the clause may vary depending on the information status of its referent (Birner and Ward, 1998). An example of this in English is ditransitive verbs, which have two variants, the to-dative (*I gave the book to the manager*) and the double-object (*I gave the manager the book*). Without a context both forms are equally acceptable, and in context native speakers may be unable to consciously decide which is more appropriate. However, the use of the forms is highly systematic and almost entirely predictable from the relative information status and the relative size of the object NPs (Snyder, 2003). In general, older, lighter NPs precede newer, heavier NPs.

Generating the appropriate ditransitive form based only on their relative frequencies is impossible, as can be seen in the behavior of the ditransitive *give* in a corpus of naturally occurring written and spoken English (Snyder, 2003).[1] Of the 552 tokens of *give* where the indirect and direct objects are full NPs,[2] 152 (27.5%) are the to-dative and 400 (72.5%) are the double object construction. Given this ratio, only the double object construction would be generated. If the distribution of relative information status and heaviness of direct and indirect objects is the same in the domain of generation as in the source corpus, then on average, the construction chosen as a surface realization will be inappropriate 3 times out of 10.

Compared to English, the evidence for the importance of word order from a free word order language like Finnish is even more striking. When word order is used to encode the information status and discourse function of NP referents, native speakers will judge the use of the wrong form infelicitous and odd, and a text incorporating several wrong forms in succession rapidly becomes incoherent (cf. Kruijff-Korbayová et al. (2002) on Czech, Russian, and Bulgarian).

Although Finnish is regarded as canonically subject-verb-object (SVO), all six permutations of these three elements are possible, and corpus studies reveal that SVO order only occurs in 56% of sentences (Hakulinen and Karlsson, 1980). Different word order variants in Finnish realize different pragmatic structurings of the conveyed information. For example, Finnish has no definite or indefinite article, and the SVO/OVS variation is used to encode the distinction between already-mentioned entities and new entities (e.g. Chesterman (1991)). OVS order typically marks the object as given, and the subject as new. SVO order is more flexible. It can be used when the subject is given, and the object is new, and also when both are old or both are new. In orders with more than one preverbal argument (SOV, OSV), as well as verb-initial orders (VOS, VSO), the initial constituent is interpreted as being contrastive (Vilkuna (1995); and others).

Because different orders have different discourse properties, use of an inappropriate order can lead to severe misunderstandings, including difficulty in interpreting NPs. For example, if a speaker uses canonical SVO order in a context where the subject is discourse-new information but the object has already been mentioned, the hearer will tend to have difficulty interpreting the NPs because OVS—not SVO—is the order that usually marks the object as discourse-old and subject as discourse-new. Psycholinguistic evidence from sentence processing experiments shows that humans are very sensitive to the given-new information carried by word order (Kaiser, 2003). Hence, it is an important factor in the quality of linguistic output of a NLG system.

Attempts to choose the appropriate word order in Finnish will encounter the same problem found with English ditransitives. Table 1 illustrates the frequency of the different word orders in a 10,000 sentence corpus used by Hakulinen and Karlsson (1980). The most frequent order is SV(X), where X is any non-subject, non-verbal constituent, and so this order should always be the one selected by a statistical algorithm. Based on the counts then, assuming that the proportion of discourse contexts is roughly similar within a domain, in only 56% of contexts will the choice of SV(X) order actually match the discourse conditions in which it is used.

Order	SV(X)	XVS	SXV	XSV	Other
N	5674	1139	60	348	2928
%	56	11	1	3	29

Table 1: Finnish word order frequency

The point here is not that statistical approaches to NLG

[1] This corpus consists of two novels, the Switchboard corpus, and a corpus of online newsgroup texts.

[2] She omits pronominal NPs because their ordering is affected by additional phonological factors related to cliticization.

are entirely flawed. Attempting to generate natural language by mimicking a corpus of naturally-occurring language may be the most practical strategy for designing robust, scalable NLG systems. However, human language is not just a system for concatenating words (or assembling trees) to create grammatical outputs. Speakers do not put constituents in a certain order simply because the words they are using to express the constituents have been frequently put in that order in the past. Constituents (and thereby words) appear in particular orders because those orders can reliably indicate the content speakers wish to communicate. Because of the lucky coincidence that statistical NLG has been primarily based on English, where the effects of word order variation are subtle, the problems with selecting a form f based only on a calculation of $P(f)$ are not obvious. It might seem as if the most frequent tree can express a given proposition adequately. However, given the English word order phenomenon shown above, a model based on $P(f)$ is problematic. Moreover, in languages like Finnish, even the generation of simple transitive clauses may result in output which is confusing for human users.

NLG must take into account not just grammaticality but contextual appropriateness, and so statistical algorithms need to be provided with an augmented representation from which to learn—not just strings or trees, but pairings of linguistic forms, contexts, and meanings. The probability we need to maximize for NLG is the probability that f is used given a meaning to be expressed and the context in which f will be used, $P(f|\textit{meaning,context})$.

4 An alternative approach

This section describes a very simple example of how a probability like $P(f|\textit{meaning,context})$ could be utilized as part of a surface realization algorithm for English ditransitives, in particular for the verb *give*. This example is only a small subset of the larger problem of surface realization, but it illustrates well the improvement in performance of using $P(f|\textit{meaning,context})$ vs. $P(f)$, when evaluated against actual corpus data.

First, the corpus from which the probabilities are being taken must be annotated with the additional meaning information conditioning the use of the form. For ditransitives, this is the information status of the indirect object NP, in particular whether it is hearer-new. Hearer-status can be quickly and reliably annotated and has been widely used in corpus-based pragmatic studies (Birner and Ward, 1998). It could be applied as an additional markup of a corpus to be used as input to a statistical generation algorithm, like the Penn Treebank, such that each NP indirect object of a ditransitive verb would be given an additional tag marking its hearer status. Here we use the corpus counts presented in Snyder (2003) for the verb *give* as our training data. Table 2 shows the frequency of the properties of hearer-newness and relative heaviness of indirect objects (IOs) and direct objects (DOs) with respect to the two ditransitive alternations.

IO STATUS		*TO*-DATIVE	DOUBLE OBJECT
Hearer-new	–	60	0
Hearer-old	IO heavier	79	31
	DO heavier	7	357
	IO=DO	6	12
	Totals	152	400

Table 2: Corpus freq. of ditransitives (Snyder, 2003)

To demonstrate the performance of an approach which counts only form, we use the equation $P(f)$ to determine the choice of double-object vs. to-dative. The relative probabilities of each order in the Snyder (2003) corpus are .725 and .275 for double object and to-dative, respectively. As such, this method will always select the double object form, yielding an error rate of 27.5% on the training data, as shown in the row labeled $P(f)$ of Table 3.

An algorithm which incorporates more information than just raw frequencies will proceed as follows: if the IO is hearer-new, generate a to-dative because the probability in the corpus of finding a to-dative given that the indirect object is hearer-new is 1 (60 out of 552 tokens). In all other cases (i.e. all other information statuses of IO and DO), the probability of finding a to-dative is now 92/400, or 18.6%, so generate a double object. This method results in 92 incorrect forms (all cases where the double object is generated instead of a to-dative), an error rate of 16.7% on the training data.

If the generation algorithm is further augmented to take into account information about the relative heaviness of the direct and indirect object NPs—possible in a system where NPs are generated separately from sentences as a whole, the error rate can be reduced even more. This algorithm will be as follows, if the IO is hearer-new, the form chosen is a to-dative. If the IO is not hearer-new, the IO and DO are compared with respect to number of syllables. If the IO is longer, generate a to-dative; if the DO is longer, generate a double object. As before, the first rule applies to the 60 tokens where the IO is hearer-new. Out of the remaining 492 tokens, 474 have IOs and DOs of different heaviness. In 357 of the 388 double objects, the DO is heavier, and in 79 of the 86 to-datives, the IO is heavier. This leaves 38 of 474 tokens not covered by the heaviness rule, along with 18 tokens where the IO and DO are equal. For these 56 cases, we generate the more probable overall form, the double object. In total then, this augmented generation rule will yield 139 to-datives (60 cases where the IO is hearer-new and 79 cases where the IO is heavier). With this algorithm, only 13 actual to-datives will be generated wrongly as double-

objects when compared to their actual form in the corpus, an error rate of only 2.4%

	DO-IO	IO-DO	Error
Actual counts	152	400	–
$P(f)$	0	552	27.5%
Hearer-status, $P(f)$	60	492	16.7%
Hearer-status, heaviness, $P(f)$	139	413	2.4%

Table 3: Error rates with respect to choice of word order

This example shows that for some arbitrary generation of a surface realization of the predicate GIVE, simply including the hearer-status of the recipient as a condition on the choice of form yields the order that matches the "gold standard" of human behavior in a meaningful way about 80% of the time vs. only 70% for an approach based on counts of trees including *give* alone. By including additional information about the relative size of the NPs, the surface realization will match the gold standard over 97% of the time, a highly human-like output.

5 Implementation & implications for NLG

The approach argued for above is one where discourse context and meaning must be taken into account when selecting a construction for NLG purposes. Admittedly, the demonstration of the error rate here is not derived from an actual system. However, functioning NLG systems have been implemented where exactly such information conditions the algorithm for choice of main clause word order (Stone et al., 2001; Kruijff-Korbayová et al., 2002). Additionally, an approach like Bangalore and Rambow's could easily be extended by annotating their corpus for hearer-status of NPs. The necessary information could also possibly be extracted automatically from a corpus like the Prague Dependency Treebank which includes discourse-level information relevant to word order. For phenomena which have not been as closely studied as English ditransitives, machine learning could be used to find correlations between context and forms in corpora which could be incorporated into statistical NLG algorithms.

The primary implication of our argument here is that counting words and trees is not enough for statistical NLG. Meaning, semantic and pragmatic, is a crucial component of natural language generation. Despite the desire to lessen the need for labeled data in statistical NLP, such data remain crucial. Efforts to create multi-level corpora which overlay semantic annotation on top of syntactic annotation, such as the Propbank (Kingsbury and Palmer, 2002), should be expanded to include annotations of pragmatic and discourse information and used in the development of statistical NLG methods. We cannot generate forms by ignoring their meaning and expect to get meaningful output. In other words, if the input to an NLG system does lack distinctions that play a crucial role in human language comprehension, the system will not be able to overcome this lack of quality and generate high-quality output.

In addition, in the effort to push the boundaries of statistical techniques, limiting the scope of research to English may give falsely promising results. If one of the primary benefits of statistical techniques is robust portability to other languages, presentation of results based on experimentation on a small subset of human languages must be accompanied by a typologically-informed examination of the assumptions underlying such experiments.

References

Bangalore, S., and O. Rambow. 2000. Exploiting a probabilistic hierarchical model for generation. In *COLING*.

Birner, B., and G. Ward. 1998. *Information status and noncanonical word order in English*. Amsterdam: John Benjamins.

Chesterman, A. 1991. *On definiteness*. Cambridge: CUP.

Gundel, J., N. Hedberg, and R. Zacharski. 1993. Cognitive status and the form of referring expressions. *Language* 69:274–307.

Hakulinen, A., and F. Karlsson. 1980. Finnish syntax in text. *Nordic Journal of Linguistics* 3:93–129.

Kaiser, E. 2003. The quest for a referent: A crosslinguistic look at reference resolution. Doctoral Dissertation, University of Pennsylvania.

Kingsbury, P., and M. Palmer. 2002. From Treebank to Propbank. In *LREC-02*. Las Palmas, Spain.

Kruijff-Korbayová, I., G. J. Kruijff, and J. Bateman. 2002. Generation of contextually appropriate word order. In *Information sharing*, 193–222. CSLI.

Langkilde, I., and K. Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *COLING-ACL*.

Prince, E. F. 1992. The ZPG letter: subjects, definiteness, and information-status. In *Discourse description*. Amsterdam: John Benjamins.

Ratnaparkhi, A. 2000. Trainable methods for surface natural language generation. In *ANLPC 6–NAACL 1*.

Snyder, K. 2003. On ditransitives. Doctoral Dissertation, University of Pennsylvania.

Stone, M., C. Doran, B. Webber, T. Bleam, and M. Palmer. 2001. Communicative-intent-based microplanning: the Sentence Planning Using Description system. Rutgers University.

Vilkuna, M. 1995. Discourse configurationality in Finnish. In *Discourse configurational languages*, ed. K. Kiss, 244–268. New York: Oxford University Press.