

A method for comparing fluency measures and its application to ITS natural language question generation

Roy Wilson

Learning Research Development Center
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, Pennsylvania 15260
rwilson@pitt.edu

Abstract

The motivation for this empirical study is to make it easier to achieve relatively more fluent dialogue. I sketch a framework for automatically assessing sentence fluency and introduce some statistical techniques that may make corpus characterization and comparisons easier. The technical approach is based on interpreting the language used in a particular situation as a sublanguage of English. Corpus-based statistical models of the English language and its putative sublanguage are constructed and two composite measures of sentence fluency are compared via four non-parametric statistical tests. With one exception, the quantitative results agree with qualitative expectations. Several additional issues are considered and directions for future work are presented.

Introduction

To the extent that teaching and tutoring are didactic in orientation, it is no surprise that much of the work in tutorial dialogue systems is oriented to question *answering* (Mittal & Moore 1995). Research on collaborative response generation, however, highlights the importance of question *asking*, especially in the conduct of clarification subdialogues (Chu-Carroll & Carberry 1998, p. 371). Questions provide one vehicle for clarification and there have been a number of attempts to formulate a “logic of questions” (Belnap & Steel 1976; Harrah 1963), including at least one consideration of its pedagogical implications (Harrah 1973). One recent study attempts to formalize the sense in which a question “arises” from a set of prior statements and questions (Wisniewski 1995, p.2). Question *asking*, then, may be an important theoretical issue in the design of tutorial dialogue systems.

Heavy reliance upon questioning is, of course, strongly associated with the “Socratic method” of teaching and tutoring. In a recent experimental study, tutors were instructed to conduct either Socratic or Didactic tutoring sessions. In the former condition, tutors were instructed to ask questions rather than give explanations: in the latter, tutors were instructed to first explain, then question. Results indicated that, indeed, Socratic tutoring differs in a statistically significant sense from Didactic tutoring in the

number of questions asked by tutors (Core, Moore, & Zinn 2002). Moreover, although the jury is still out, other recent empirical work suggests that Socratic learning is more effective than Didactic learning with respect to student learning gains (Rose *et al.* 2000). Hence, there is empirical support for the pedagogic importance of, as well as theoretical grounding for, question asking by an intelligent tutoring system (ITS).

This study presents an analysis of a set of wh-interrogatives derived from transcripts of human-human tutoring of qualitative physics. This set samples the kind of subject matter and pedagogical expertise that natural language generator used in the Why-Atlas ITS is intended to (*cautiously* (Reiter 2002)) approximate. The use of such a set of wh-interrogatives reflects the view that corpora may capture the high quality of language used by domain experts (Oh & Rudnicky 2002).

Given the intuition that one corpus may better capture expert language use than another, one might ask, “Why bother, then, to compare measures of fluency that are derived from qualitatively different corpora?” In an information theoretic sense, it might be said, there is no gain. Some have asserted, however, that the *quantitative* characterization and comparison of corpora is useful in its own right (Kilgariff 2001).

I ask the following empirical research question: “In a statistically significant sense, do two (composite) measures of sentence fluency differ and do they differ in the extent to which they agree with human judgment?” The above question is answered by the novel application of non-parametric statistical techniques to the produced by these measures. I now provide the background and motivation for the study and a description of the corpora and language models that make it possible to address the above empirical research question.

Background

The Why-Atlas ITS encourages students to write, and provide detailed support for, their answers to qualitative physics problems. It does so by engaging the student in dialogue. A set of (computer-mediated) human-human dialogues between expert tutors and students were logged as part of a pilot study of the Why-Atlas ITS. The following excerpts were drawn from a typical interaction between an expert human tutor and a student.

Problem: Suppose you are running in a straight line at constant speed. You throw a pumpkin straight up. Where will it land? Explain.

Student: The pumpkin should fall right beside you if you continue at the constant speed in a straight line.

Tutor: Can you explain a little about the vertical motion of the pumpkin?

Student: The pumpkin will have a vertical motion and a horizontal motion which both act to make the pumpkin land right beside you.

...

Tutor: Why should it land beside you and not exactly in your hand?

Student: I guess I am wrong and the pumpkin should fall a little to the side and not in your hand.

...

Tutor: What can you say about the horizontal speed of the pumpkin in relation to your own speed?

Student: they are the same

Tutor: Good. Now you are ready to revise your essay.

As in the above example, many tutor responses in the Why-Atlas pilot evaluation corpus are wh-interrogative phrases (Trotta 2000, p. 38) with a single wh-word head such as *what*, *when*, *where*, *why* or *how*. This is not surprising, since questioning is often used by tutors to encourage explanation, direct students, and request clarification.

The Why-Atlas ITS uses the RealPro (Lavoie & Rambow 1997) surface realizer to generate clarification requests, principally in the form of wh-interrogatives. The sole input to RealPro is a Deep Syntactic Structure (DSYNTS) (Mel'čuk 1988). Given a request from the dialog manager to realize a clarification question, the current clarification module, which has been implemented and tested, employs templates (versus canned text) to create the DSYNTSs needed to generate a wh-interrogative.

Motivation for the work

As noted by a number of investigators (Oh & Rudnicky 2002), templates are costly to develop and maintain. Template-driven, DSYNTS-based, generation requires (roughly) the developer to specify a new DSYNTS for each (parameterized) class of utterance to be generated. In practice, this often means taking a previously developed DSYNTS and modifying it. Although experience and knowledge reduce the time needed to create new templates from old or from scratch, template creation is a costly process with considerable amounts of trial-and-error search. The ability to produce fluent dialog in a template-based generation system is practically limited by the costs associated with template creation.

It would be desirable, therefore, to have a computer program aid in constructing DSYNTSs. Creating a machine learning program that can create a DSYNTS based on an arbitrary criterion is a difficult, if not impossible, problem. I simplify the problem in two ways. First, I restrict the scope of the computational task to transforming one DSYNTS into

another such that the texts they produce are roughly synonymous. Second, I further restrict the computational task to that of learning to paraphrase wh-interrogatives. By restricting the learning task, it may be possible to create a machine learning program capable of assisting with template creation.

A machine learning problem such as the one described above is well-posed only if it identifies three features: the task involved; the source of experience to be acquired by the machine learning program; and a measure of task performance to be improved (Mitchell 1997, p. 2). The task of the paraphrasing system mentioned above would be to learn an optimal strategy (Mitchell 1997, p. 312) for paraphrasing based on a set of more elementary lexical and syntactic "moves". The machine learning program would acquire experience by exploring a space of DSYNTSs via sequences of elementary paraphrasing moves. Obviously, a program to learn an optimal strategy for generating fluent paraphrases of wh-interrogatives requires a measure of fluency as its performance measure. So, the limited problem of machine learning to paraphrase wh-interrogatives generated by DSYNTSs is well-posed only if it is possible to formulate a measure of sentence fluency.

Although the goal of creating a machine learning program capable of generating paraphrases informs this study, its purpose is simply to offer empirical evidence concerning the well-posedness of the restricted machine learning problem described above. In particular, this study examines two (composite) measures of sentence fluency in order to assess their potential suitability as measures of machine learning performance. Neither does this study offer a fluency measure to be used directly in generation: Given the simplicity of the models on which the fluency measures are based, as described below, the fluency measures are unlikely to have sufficient representational power and generalization capability to be useful in direct generation. Rather, the kind of fluency measures examined here are only intended to provide guidance to a machine learning program. If this and other quantitative studies suggest that no adequate measures of sentence fluency are available, then the machine learning problem as described above is not well-posed and should, perhaps, be set aside.¹

Statistical language models and their corpora

Recently, researchers investigating large-scale, domain-independent, generation combined statistics from a bi-gram statistical language model (BSLM) and scores produced by a statistical parser (ESLM, for Charniak's maximum-entropy-inspired parser) to compare the fluency of paired texts (Daumè *et al.* 2002, p. 15)². The BSLM and the ESLM both generate a real number for each input sentence they process. I take each such number as a measure of the

¹A fuller examination of the machine learning problem is underway and is being described in a technical report entitled "(Machine) Learning to paraphrase" that is in preparation.

²Although it might be very fruitful to combine measures in order to capture different aspects of fluency, this was not the primary intent of Daumè and colleagues.

fluency of the input sentence in relation to the statistical language model that produces it and the corpus used to train that model. I now describe each model.

The BSLM in this study was trained on a corpus of wh-interrogatives obtained from qualitative physics tutoring sessions, hereafter designated as the QPT corpus. The BSLM uses bi-gram statistics to calculate the fluency of *S*, where *S* is a wh-interrogative in the QPT corpus. A bi-gram is simply an ordered pair of words and the likelihood of a bi-gram in a language (or sublanguage) is estimated by its likelihood in a corpus (or subcorpus). The likelihood of a sentence *S* is calculated as the product of the likelihoods of the bi-grams associated with each word pair in the sentence. The composite measure formed from the BSLM and the QPT corpus is hereafter referred to as the BSLM-QPT measure. The BSLM-QPT measure of the fluency of *S* is defined as the likelihood of *S*.

The ESLM was trained by Charniak on the Wall Street Journal corpus. Given a wh-interrogative *S*, Charniak's parser produces both a parse of *S* and the probability of that parse. Following (Daumè *et al.* 2002), I take the probability of a parse of *S* as a measure of the fluency of *S* in relation to the ESLM and the Wall Street Journal (WSJ) corpus. The composite measure formed from the ESLM and the WSJ corpus is hereafter referred to as the ESLM-WSJ measure.

Because each of the two measures under consideration are composites formed from a statistical language model and a training corpus, no claims can be made concerning the relative superiority of the two *statistical models* with respect to the qualitative physics tutoring sublanguage. Instead, this paper describes a set of simple comparisons involving the two (composite) measures of wh-interrogative fluency. Possible steps toward disambiguating model effects and corpus effects are discussed in the last section of the paper.

Constructing the statistical data

How do the BSLM-QPT and ESLM-WSJ measures differ and which measure better reflects human judgments of fluency? One way of answering this question in a quantitative sense is to construct several statistical data sets and test a set of hypotheses pertaining to them. In this section, we describe the first step.

Constructing the QPT corpus

In order to model the sublanguage of qualitative physics tutoring, I constructed a target corpus (hereafter designated as the QPT corpus) for training the BSLM (Reiter & Dale 2000, pp. 30-36). From a collection of 41 human-human student-tutor dialogs obtained during a Why-Atlas ITS pilot evaluation, the QPT corpus was extracted in two stages. In the first stage, just those wh-interrogatives initiated by the tutor were extracted, yielding a set of approximately 4467 elements, consisting of 63,339 words. When necessary, wh-interrogatives such as "Now why should it land beside you and not exactly in your hand?" were rephrased to have the wh-word in the initial position of the wh-interrogative. Spelling errors were corrected and pronominal anaphora

(such as 'it') were resolved. In the second stage, wh-interrogatives beginning with the words *what*, *when*, *where*, *why* and *how* were extracted from the first set, yielding a set of 282 wh-interrogatives consisting of approximately 3200 words. The QPT corpus was then used to train the BSLM and thus to construct the BSLM-QPT measure.

Defining the BSLM-QPT fluency measure

From the QPT corpus, the frequency of each bi-gram was tabulated. The logarithm of the likelihood ratio statistic (LLR) was then calculated for each bi-gram using (Resnik 2001). Although large-sample statistics are sometimes used to advantage with small corpora (Fisher & Riloff 1992), the logarithm of the likelihood ratio does a better job capturing collocations in the corpus than tests based on the normal distribution (Dunning 1993). After the BSLM was trained on the QPT corpus, the BSLM-QPT fluency measure was defined as follows.

The BSLM-QPT fluency measure of a wh-interrogative *S* was defined as the sum of the LLRs of the bi-grams that constitute *S* (using the formulation of (Langkilde & Knight 1998)). If a bi-gram in *S* does not appear in the QPT corpus, it contributes nothing to the total LLR for a wh-interrogative *S* in which it does appear. Any set of wh-interrogatives can be ranked by the magnitude of the BSLM-QPT fluency measure assigned to each element of the set. Note that the LLR was not used to test hypotheses, but was used to assign numeric values to bi-grams.

Constructing the test corpus

In (Daumè *et al.* 2002), a set of handcrafted "hope sentences" were constructed, each described as more fluent than the automatically generated text *S* with which it was compared. The purpose of constructing a "hope sentence" (designated PS) was (in effect) to test the ability of various classifiers to select the more fluent member of each pair (*S*, PS) of texts. I adopt a similar sentence construction strategy, not to build a classifier, but to compare the BSLM-QPT and the ESLM-WSJ fluency measures. Of course, neither measure of fluency takes context into account, a limitation discussed in the final section.

For each wh-interrogative *S* in the QPT corpus, I attempted to construct two paraphrases, one of type MF (for "more fluent") and one of type LF (for "less fluent"). Each MF paraphrase of *S* was obtained by correcting the spelling/grammatical errors and minor lexical and syntactic dysfluencies of *S* (if any). Each LF paraphrase of *S* was obtained by deleting the initial word in a pedagogically important bi-gram of *S*. To illustrate these differences, consider the following example.

S: When an object is dropped from a height, what is it's vertical velocity just at the moment of release?

MF: When an object is dropped, what is it's vertical velocity just at the moment of release?

LF: When an object is dropped from a height, what is it's velocity just at the moment of release?

As illustrated above, the difference between an MF paraphrase of *S* and *S* is primarily syntactic, whereas the difference between an LF paraphrase of *S* and *S* is primar-

ily lexical. This particular way of creating paraphrases was chosen for simplicity of implementation and to highlight the expected strengths and weakness of each fluency measure.³ The resulting corpus, hereafter designated as the TST (for test) corpus, contained 147 paraphrases of type MF or LF, each derived from one wh-interrogative S in the QPT corpus.

Comparing the fluency measures to human judgment

As described above, each element in the TST corpus corresponds to a wh-interrogative in the QPT corpus. Each of the 147 such correspondences represents a human judgment of the form “S is more fluent than LF” or “MF is more fluent than S”. The set of 147 pairs of the form (S, LF) or (MF, S) provides a baseline for comparing the BSLM-QPT and the ESLM-WSJ fluency measures.

Suppose, for definiteness, that we are given a pair of the form (S, LF), where S is an element of QPT, LF is a corresponding element of TST, and that the measure of interest is the BSLM-QPT. The BSLM-QPT measure can be applied to both members of the pair, yielding a pair of real numbers designated in functional form by (BSLM-QPT(S), BSLM-QPT(LF)). If $\text{BSLM-QPT}(S) > \text{BSLM-QPT}(LF)$, then the fluency ordering produced by the BSLM-QPT measure agrees with the ordering produced by human judgment. Similar logic applies to the ESLM-WSJ measure and baseline pairs of the form (MF, S).

Constructing the statistical data sets

The distributional properties of the data produced by the BSLM-QPT and the ESLM-WSJ measures are unknown. To avoid ungrounded distributional assumptions, such as normality, all statistical analyses are based on the ranks associated with the data rather than the data themselves. The distributional properties of the ranks, which in the simplest case run from 1 to N where N is the number of elements to be ranked, are known (Conover 1980).

Consider the following, highly simplified, example. Suppose we have only three wh-interrogatives to consider: S_1 , S_2 , and S_3 . Suppose the BSLM-QPT measure produces the following three real numbers: 3.5, 16.2, and 15.0. This means that the BSLM-QPT ranks S_1 as the most fluent and S_2 as the least fluent of the three wh-interrogatives, so that the ranking produced by the BSLM-QPT is 1, 3, 2. Suppose now that, for the same three wh-interrogatives, the ESLM-WSJ measure produces the following three real numbers: 103.5, 16.2, and 175.0. This means that the ESLM-WSJ produces the following ranking: 2, 1, 3. The ranking produced by a fluency measure constitutes a statistical data set.

To compare the BSLM-QPT and ESLM-WSJ measures, I compare the statistical data sets they produce, asking whether these data sets differ in a statistically significant sense. If either set of real numbers contains ties, the average of the ranks that would have been assigned without ties

³Note that, in this instance, the human tutor did not communicate via LF, which might be the anaphoric equivalent of S.

is computed and assigned to each tied value. It is possible to give a nonparametric statistical analysis of the ranked data to determine whether the BSLM-QPT and the ESLM-WSJ differ quantitatively and whether they differ in the extent to which they agree with human judgment. Although corpus-based measures of fluency are not new, as noted earlier, the nonparametric statistical analysis of the data produced by such measures is, as far I know, new.

Analyzing the statistical data

Hypothesis 1

I expect systematic differences in how the BSLM-QPT and the ESLM-WSJ rank the wh-interrogatives in the QPT corpus because each statistical language model is trained on a different corpus. On the other hand, each corpus is a subset of the English language, so I expect substantial agreement in how the BSLM-QPT and the ESLM-WSJ rank the wh-interrogatives in the QPT corpus. Hence, I expect a “substantial” positive correlation between the fluency ranking produced by the ESLM-WSJ and that produced by the BSLM-QPT, an expectation tested via the following statistical null hypothesis.

On the 282 wh-interrogatives in the QPT corpus, the fluency rankings produced by the BSLM-QPT and the ESLM-WSJ are uncorrelated or negatively correlated.

Spearman’s bivariate correlation coefficient ρ is calculated as follows. The rankings produced by the BSLM-QPT can be thought of as the first variable, and the rankings produced by the ESLM-QPT can be thought of as the second variable. A Pearson bivariate correlation calculated based on the two variables yields Spearman’s ρ .

With $N = 282$, a Spearman rank correlation of $\rho = 0.31$ was obtained. Hence, the two measures are systematically related. The attained significance of the hypothesis is 0.0001 and so the null hypothesis is rejected: it is more likely that the fluency rankings are positively correlated.

Is ρ “substantial”? A Pearson correlation of 0.31 is associated with an effect size of approximately 0.65. According to Cohen’s standard, an effect size of 0.65 is a “medium” effect. I take this as evidence that the magnitude of ρ is substantial.

Additional analysis shows differences in how particular classes of wh-interrogatives are ranked by the BSLM-QPT and the ESLM-WSJ. For example, wh-interrogatives from the QPT corpus that are headed by wh-words other than *what* are ranked higher by the BSLM-QPT than the ESLM-WSJ. Although these differences appear to be statistically significant, additional research is needed.

Hypothesis 2

Recall that the TST corpus consists of wh-interrogatives constructed to be either more fluent (MF) or less fluent (LF) than some wh-interrogative S in the QPT corpus. There are 147 such pairings. Consider one such pair of wh-interrogatives (S, MF). When applied to each element of the pair, each measure produces a corresponding pair of real

numbers (v_1, v_2) . If $v_1 < v_2$, then the fluency ordering produced by the measure agrees with the fluency ordering constructed by the human judge, otherwise it does not.

In looking at how corpus variation affects parser performance, Gildea notes strong corpus effects and that lexical bigram statistics appear to be corpus-specific (Gildea 2001). Because the BSLM is trained on the QPT corpus, I expect the BSLM-QPT measure to produce pairwise fluency orderings that show a higher level of agreement with human judgment than the pairwise fluency orderings produced by the ESLM-WSJ, an expectation that is tested via the following statistical null hypothesis.

When comparing each of the 147 wh-interrogatives in the TST corpus with its associated wh-interrogative in QPT corpus, the BSLM-QPT measure shows a level of agreement with human judgment that is less than or equal to that shown by the ESLM-WSJ measure.

Hypothesis 2 (as well as 3 and 4) is tested using Cochran's nonparametric test (Conover 1980, p. 199). Cochran's test⁴ can be thought of as a nonparametric one-way analysis of variance. McNemar's test is ordinarily used to analyze a 2-by-2 table that gives the number of times two measures agree with each other. Although McNemar's test is equivalent to a restricted form of Cochran's test, Cochran's test is applicable when comparing more than two measures, whereas McNemar's test is not. For these reasons, the analyses carried out in connection with hypotheses 2 through 4 are based on Cochran's test.

The null hypothesis is rejected: Cochran's test yields a test statistic of 17.66 that is significant at the 0.0001 level. Overall, the BSLM-QPT measure produces a fluency ordering that is in significantly greater agreement with a human judge of relative fluency than the fluency ordering produced by the ESLM-WSJ measure. With respect to the 147 pairs of wh-interrogatives formed from the TST corpus and the QPT corpus, the BSLM-QPT agreed with human judgment of relative fluency on 46 percent, and the ESLM-WSJ agreed with human judgment on 24 percent, of the pairs.

Hypothesis 3

Recall that wh-interrogatives of type MF in the TST corpus are constructed by minor grammatical or lexical improvements of a wh-interrogative S in the QPT corpus. Since the ESLM-WSJ is based on a broader model of the English language, it should be more sensitive to such changes than the linguistically narrower BSLM-QPT. When comparing the fluencies of an MF wh-interrogative and S, then, the ESLM-WSJ measure ought to track human judgment better than the BSLM-QPT measure, an expectation tested via the following statistical null hypothesis.

⁴Readers may wonder why Cohen's kappa κ is not used to assess the level of agreement between the orderings produced by each fluency measure and the human rater, so that the resulting kappas could then be compared. It can be illustrated by example and proved using elementary algebra (interested readers should contact the author) that, when comparing a single rater to a standard, *any* κ that is computed will necessarily be less than or equal to zero. In this situation, the κ statistic is not a good measure of agreement.

When comparing the fluency of each wh-interrogative of type MF in the TST corpus with the wh-interrogative S in the QPT corpus from which it was derived, the ESLM-WSJ shows a level of agreement with human judgment that is less than or equal to that shown by the BSLM-QPT.

Cochran's test yields a test statistic of 0.04 so there is insufficient statistical evidence to reject the null hypothesis. In other words, there is no reason to believe that the ESLM-WSJ tracks human relative fluency judgment any better than the BSLM-QPT. In fact, both the BSLM-QPT and the ESLM-WSJ measures show only 28 percent agreement with the relative fluency ordering of the human judge.

Hypothesis 4

Recall that a wh-interrogative of type LF in the TST corpus is obtained from a wh-interrogative S in the QPT corpus by replacing a substantively important word-pair in S with a less important one. Since the BSLM is trained on the QPT corpus, it should be more sensitive to the absence of language characteristic of domain expertise. When comparing the fluencies of S and LF, then, the BSLM-QPT measure ought to track human judgment better than the ESLM-WSJ measure, an expectation tested via the following statistical null hypothesis.

When comparing the fluency of each wh-interrogative of type LF in the TST corpus with the wh-interrogative S in the QPT corpus from which it was derived, the BSLM-QPT shows a level of agreement with human judgment that is less than or equal to that shown by the ESLM-WSJ.

The null hypothesis is rejected: Cochran's test yields a test statistic of 31.11 that is significant at the 0.0001 level. The BSLM-QPT measure shows significantly greater agreement with human judgment than the ESLM-WSJ measure. The BSLM-QPT measure agrees with the human rater on 97 percent, whereas the ESLM-WSJ agrees on only 11 percent, of the 38 (S, LF) pairs.

Related work

On sublanguage

Students are encouraged to learn not only physics concepts, but the language of physics as well. Not surprisingly, the vocabulary of human-human qualitative physics tutoring sessions is dominated by terms from introductory mechanics such as velocity, acceleration, displacement, and force. Roughly speaking, human tutors and the Why-Atlas ITS employ the language of "qualitative physics tutoring" (QPT).

Although an adequate empirical definition of the term may be lacking (Kittredge 1982, p. 110), it has been said that "sublanguage" denotes a theoretical construct (Lehrberger 1986, p. 22). Among the empirical indicators associated with it are: a semantically limited domain of discourse; shared habits of word usage; the fact that "[t]he specific word co-occurrences which obtain in a sublanguage are a

reflection of the special domain and its organization” (Kittredge 1982, p. 119). For example, the pair “horizontal velocity” is more common in logged QPT sessions than “the velocity”, even though the latter phrase is often a perfectly acceptable (anaphoric) substitute. This is so, I conjecture, because the first pair expresses a domain concept that is important for students to understand and about which they often have misconceptions or a lack of clarity. Since logged QPT sessions exhibit the indicators associated with the sublanguage phenomena, it seems reasonable to think of the language used in QPT sessions as a sublanguage of English.

The originators of the “sublanguage” concept used formal grammar to model sublanguage. Others who have drawn on the idea of “sublanguage” (McKeown, Kukich, & Shaw 1994), however, have modeled the concept in a looser sense. Rather than formulate a grammar, I develop a bi-gram statistical model of the QPT sublanguage based on the co-occurrence of words in a corpus.

Can a bi-gram statistical language model adequately represent a sublanguage? It has been claimed that collocations, defined as arbitrary and recurrent word combinations, can be used in generation in a straightforward manner (McKeown & Radev 2000). Although bi-gram statistics alone are not sufficient to distinguish between idioms, collocations, and free word combinations, bi-gram statistics *have* been used with other techniques to extract collocations from a large corpus (Smadja & McKeown 1992). If collocations can be used straightforwardly in generation and bi-gram statistics are helpful in identifying collocations, there is reason to hope that bi-gram statistics may provide sufficient information to distinguish more and less fluent texts.

On sentence fluency

Early NLG workers argued that the absence of a general theory, and the domain-specificity, of natural language processing made it necessary to design domain-dependent (rather than domain-independent) systems on the basis of a knowledge-engineering approach (Kukich 1983, pp. 19-20). Accordingly, an early approach to the issue of generation fluency sought to identify the skills and defects exhibited by a system and argued that greater fluency would follow from more effective identification and integration of the knowledge processes that drive the generation of fluent text (Kukich 1988). A knowledge-acquisition bottleneck has, however, led an increasing number of NLG researchers to turn to statistical or hybrid representations of knowledge.

A more recent analysis of generation fluency proposed the use of n-grams for large scale, domain independent, sentence generation (Knight & Hatzivassiloglou 1995). Nitrogen, which was trained on the Wall Street Journal corpus, is one such system. Given a set of candidate text realizations (over-)generated from a symbolic form, Nitrogen used bi-gram statistics to identify the most fluent sentence realized (Langkilde & Knight 1998). Recent work by (Oh & Rudnicky 2002) adopts multiple n-gram models for corpus-based, spoken language generation in the travel domain. Whereas Nitrogen is aimed at domain-independent generation, this study targets domain-dependent generation. Oh and Rudnicky use multiple statistical language models for

speech generation in a domain that may be more limited than that of qualitative physics tutoring: this study uses a single bi-gram language model to capture both qualitative physics knowledge and domain communication knowledge (Kittredge, Korelsky, & Rambow 1991).

The fluency of generated text is difficult to assess (Kukich 1988). Indeed, most workers have shied away from attempting to formulate a positive concept of fluency, relying instead on the ability to recognize dysfluencies. The assumption made here is that the lexical regularities characteristic of the QPT sublanguage give reason to hope that bi-gram statistics based on a small corpus (Fisher & Riloff 1992) can supply enough representational power and generalization ability so that the BSLM-QPT measure, or something like it, is an adequate performance measure for the machine learning problem that motivated this study.

On questioning

Research on generation in clarification dialogues (Churroll & Carberry 1998; Cohen, Schmidt, & van Beek 1994) highlights the importance of question *asking*. The work done here does not, however, occur within the context of plan recognition, except perhaps in a very broad sense. Recent work has examined the questioning strategies used by human tutors of qualitative physics, creating a taxonomy of question types (Jordan & Siler 2002). Unlike that work, this study is based solely on lexico-syntactic form. Finally, I do not attempt to tackle the pragmatic issues involved in questioning (Wisniewski 1995, p.2) and in the natural language generation (Hovy 1988) of the wh-interrogatives that realize questions.

Summary and future directions

I have addressed the following empirical research question: “In a statistically significant sense, do the BSLM-QPT and ESLM-WSJ measures of fluency differ and do they differ in the extent to which they agree with human judgment?” In the process of obtaining these results I have sketched a framework for automatically assessing sentence fluency in a sublanguage context and introduced some statistical techniques that may make corpus characterization and comparison easier. As always, several issues deserve further consideration.

In terms of the sentence planning architecture described in (Reiter & Dale 2000, pp. 122-123), DSYNTS-based clarification generation in the Why-Atlas ITS currently focuses on deciding which words and syntactic structures to use to express a question. It does yet not, however, take into account the role of context: in particular, it does not address referring expression generation, a factor important to any consideration of sentence fluency in the context of dialog. In future generation work, I intend to selectively integrate the stochastic approach to generation taken by Oh and Rudnicky (Oh & Rudnicky 2002) (which does address the generation of referring expressions) and compare it with a template-based implementation of referring expression generation. I also intend to explore the possible aggregation of DSYNTSs along the lines taken in (Walker, Rambow, & Rogati 2002).

This study bears the mark of resource limitation. Having multiple persons compare the fluency of the wh-interrogative pairs might be better than relying on a single human being. Using more of the original corpus would also be desirable as it would increase the sample size, which might increase the sensitivity of the BSLM-QPT measure and perhaps reveal areas where the ESLM-WSJ measure outperforms it. Although removing these resource limitations is clearly desirable, it is not clear to what extent the results obtained were influenced by those limitations.

It is desirable to understand whether the relatively poor performance of the ESLM-WSJ is due to the fact that it was trained on the Wall Street Journal corpus, to the underlying statistical model, or to some possibly Byzantine interaction of corpus and model. One step toward separating corpus effects from model effects would be to train both the BSLM and the ESLM on the subset of 700 randomly selected sentences from the Wall Street Journal corpus recently made available (XeroxPARC 2002) and then test each on the corpora constructed from the QPT sessions. Another possibility would be to train Charniak's latest (November 2001) parser on the QPT corpus after converting that corpus into Penn Treebank form. These two steps would provide greater understanding about the interaction between statistical language models and their corpora in the creation of composite fluency measures.

Although it seemed reasonable that the ESLM-WSJ measure would outperform the BSLM-QPT measure when comparing a wh-interrogative S with a MF wh-interrogative, that outcome did not occur. It may be better to consider the task of comparing the fluency of two wh-interrogatives (or, more generally, texts) as a classification task (as done, informally, in (Daumè *et al.* 2002, p. 14) and in a more formal way in (Bangalore & Rambow 2000)) based on multiple features. One advantage of Cochran's test is that it is well suited for comparing different feature sets that each result in binary outcomes.

Finally, and more broadly, I intend to consider the pragmatics of question asking. First, I intend to make use of the annotated data on questions collected by (Jordan & Siler 2002). Second, I intend to examine such data in the light of the possibility of formalizing the inferential process by which questions "arise" from sets of declarative statements (Wisniewski 1995). By doing so, I hope to generate wh-interrogatives that better satisfy the dialogue goals of clarification.

Acknowledgments

This research was supported by MURI grant N00014-00-1-0600 from ONR Cognitive Science. I thank the entire NLT team for helping create an environment supporting such research. In particular, I thank Kurt VanLehn and Pamela Jordan for: the opportunity to do the research; helping improve this paper; and, most of all, prompting me to make more explicit the assumptions that shaped the work. Thanks are also due to the anonymous reviewers.

References

- Bangalore, S., and Rambow, O. 2000. Evaluation metrics for generation. citeseer.nj.nec.com/bangalore00evaluation.html.
- Belnap, N. D., and Steel, T. B. 1976. *The logic of questions and answers*. Yale University Press.
- Chu-Carroll, J., and Carberry, S. 1998. Collaborative response generation in planning dialogues. *Computational Linguistics* 24(3):355–400.
- Cohen, R.; Schmidt, K.; and van Beek, P. 1994. A framework for soliciting clarification from users during plan recognition. In *Proceedings of the Fourth International Conference on User Modeling*, 355–400.
- Conover, W. J. 1980. *Practical Nonparametric Statistics*. John Wiley and Sons.
- Core, M. G.; Moore, J. D.; and Zinn, C. 2002. Initiative in tutorial dialogue. In *Proceedings of the ITS 2002 Workshop on Empirical Methods for Tutorial Dialogue Systems*, 46–55.
- Daumè, H. I.; Knight, K.; Langkilde-Geary, I.; Marcu, D.; and Yamada, K. 2002. The importance of lexicalized syntax models for natural language generation tasks. In Rambow, O., and Stone, M., eds., *Second International Natural Language Generation Conference*, 9–16. Harriman, NY: Association for Computational Linguistics.
- Dunning, T. E. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1):61–74.
- Fisher, D., and Riloff, E. 1992. Applying statistical methods to small corpora: Benefitting from a limited domain. In *Working Notes of AAAI Fall Symposium Series*, 47–53. Cambridge, MA: AAAI Press.
- Gildea, D. 2001. Corpus variation and parser performance. <http://citeseer.nj.nec.com/gildea01corpus.html>.
- Harrah, D. 1963. *Communication: A logical model*. Cambridge, MA: MIT Press.
- Harrah, D. 1973. The logic of questions and its relevance to instructional science. *Instructional Science* 1:447–467.
- Hovy, E. H. 1988. *Generating natural language under pragmatic constraints*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.
- Jordan, P. W., and Siler, S. 2002. Student initiative and questioning strategies in computer-mediated human tutoring dialogues. In *Proceedings of the ITS 2002 Workshop on Empirical Methods for Tutorial Dialogue Systems*, 39–45.
- Kilgarriff, A. 2001. Comparing corpora. *International Journal of Corpus Linguistics* 6(1):97–133.
- Kittredge, R.; Korelsky, T.; and Rambow, O. 1991. On the need for domain communication knowledge. *Computational Intelligence* 7(4):305–314.
- Kittredge, R. 1982. Variation and homogeneity of sublanguages. In Kittredge, R. I., and Lehrberger, J., eds., *Sublanguage: Studies of Language in Restricted Semantic Domains*. de Gruyter.

- Knight, K., and Hatzivassiloglou, V. 1995. Two-level, many-paths generation. In *Proceedings of the Thirtythird Conference of the Association of Computational Linguistics (ACL-95)*, 252–260. Boston, MA: Association for Computational Linguistics.
- Kukich, K. 1983. *Knowledge-based report generation: a knowledge engineering approach to natural language report generation*. Ph.D. Dissertation, University of Pittsburgh.
- Kukich, K. 1988. Fluency in natural language reports. In McDonald, D. D., and Bolc, L., eds., *Natural Language Generation Systems*. New York, NY: Springer-Verlag.
- Langkilde, I., and Knight, K. 1998. The practical value of N-grams in generation. In Hovy, E., ed., *Proceedings of the Ninth International Workshop on Natural Language Generation*. New Brunswick, New Jersey: Association for Computational Linguistics. 248–255.
- Lavoie, B., and Rambow, O. 1997. A fast and portable realizer for text generation systems. In *Proceedings of the Fifth Conference on Applied Natural Language Processing Chapter of the Association for Computational Linguistics*, 265–268.
- Lehrberger, J. 1986. Sublanguage analysis. In Grishman, R., and Kittredge, R., eds., *Analyzing language in restricted domains: Sublanguage description and processing*. Lawrence Erlbaum Associates, Publishers.
- McKeown, K. R., and Radev, D. R. 2000. Collocations. <http://citeseer.nj.nec.com/mckeown00collocations.html>.
- McKeown, K.; Kukich, K.; and Shaw, J. 1994. Practical issues in automatic documentation generation. In *Proceedings of the Fourth ACL Conference on Applied Natural Language Processing (13–15 October 1994, Stuttgart)*.
- Mel'čuk, I. A. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press.
- Mitchell, T. M. 1997. *Machine Learning*. McGraw-Hill.
- Mittal, V. O., and Moore, J. D. 1995. Dynamic generation of follow up question menus: facilitating interactive natural language dialogues. In *Human Factors in Computing Systems: CHI '95 Conference Proceedings*, 90–97.
- Oh, A. H., and Rudnicky, A. I. 2002. Stochastic natural language generation for spoken dialog systems. 16:387–407.
- Reiter, E., and Dale, R. 2000. *Building Natural Language Generation Systems*. Cambridge, UK: Cambridge University Press.
- Reiter, E. 2002. Should corpora texts be gold standards for nlg? In Rambow, O., and Stone, M., eds., *Second International Natural Language Generation Conference*, 99–104. Harriman, NY: Association for Computational Linguistics.
- Resnik, P. 2001. C program to calculate simple lr. <http://www.cs.pitt.edu/~litman/courses/CS3730/hws/ngrams.html>.
- Rose, C. P.; Moore, J. D.; VanLehn, K.; and Allbritton, D. 2000. A comparative evaluation of socratic versus didactic tutoring.
- Smadja, F., and McKeown, K. R. 1992. Using collocations for language generation. 7(4):229–239.
- Trotta, J. 2000. *Wh-clauses in English: Aspects of theory and description*. Atlanta, GA: Rodopi.
- Walker, M. A.; Rambow, O. C.; and Rogati, M. 2002. Training a sentence planner for spoken dialogue using boosting. *Computer Speech and Language*.
- Wisniewski, A. 1995. *The posing of questions: Logical foundations of erotetic inferences*. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- XeroxPARC. 2002. The parc 700 dependency bank. <http://www2.parc.com/istl/groups/nltt/fsbank/>.