Towards a corpus based heuristic for assessing communal common ground

Roman Kutlak, Kees van Deemter, Chris Mellish

April 16, 2012

Keywords: Common Ground; Estimation; Evaluation with human subjects; Web as corpus

1 Introduction

In this article we present a corpus-based heuristic for determining what facts are likely to be known by most or all the members of a given community. We believe that a heuristic of this kind could be useful in a large range of applications, computational and otherwise. For example, the heuristic could help a Natural Language Generation system to avoid stating facts that are widely known, thus diminishing the chance that the reader will get bored or irritated. Another example is a system that uses formal argumentation theory to present arguments to people; such a system could be improved by linking arguments to facts that are widely known. For example, to argue that butter is unhealthy, it might suffice to say "butter increases cholesterol", because it is widely known (or believed) that cholesterol can cause heart disease.

Our goal is to use the proposed heuristic in determining what facts to mention in a description of a famous person.

2 Common Ground

It is widely accepted in psycholinguistics that speakers aim to provide enough information for hearers but not more than necessary as this might mislead the hearers (Grice, 1975). Systems that generate natural language (e.g., FoG, Newsblaster, BabyTalk) would do well to abide by these rules in order to produce output that hearers expect. A crucial task in deciding what information to provide is to estimate what information hearers already possess. The knowledge that is shared amongst people is called common ground. Human speakers are good judges of what information is in common ground (Jucks, Becker, & Bromme, 2008; Nickerson, Baddeley, & Freeman, 1987) and use this ability frequently in communication.

The psycholinguist Herbert Clark (1996) distinguishes two different kinds of common ground: personal common ground and communal common ground. In this article we focus on communal common ground and on estimating what facts are likely to be known by a certain community. As an example of communal common ground, consider the residents of London. As they travel through the city they see Big Ben. They have no reason to suspect that others did not see it, and so they can take this as an indication that the presence of Big Ben is in common ground of the community of London residents. In other words, they can talk about Big Ben as if every Londoner knew about it.

3 Estimating Communal Common Ground

We used Herbert Clark's intuitive way of defining communal common ground and hypothesised that corpora related to the community (e.g., a collection of articles available on the Internet) can

be used as a basis for estimating communal common ground. We assumed that facts mentioned frequently in the corpora are likely to be known by the community members due to frequent exposure to these facts.

We used several corpus-based metrics of estimating the score of each fact. The simplest metric counted the number of articles that contained the fact (Frequency metric). More sophisticated metrics split each fact in two parts, the name of the person n and the statement about the person p (e.g., Albert Einstein; was a physicist), and used collocation measures to calculate the score. We compared three collocation measures: Probability of name given a fact $P(n|p)$, probability of fact given name $P(p|n)$ and a version of point-wise mutual information (PMI) (Fano, 1961) calculated as $\sqrt{count(n,p)} * log_2 \frac{P(n,p)}{P(n)P(p)}$. PMI relies on the fact that if n and p appear together more often than by a chance, the value of the fraction $\frac{P(n,p)}{P(n)P(p)}$ is bigger than 1.0. The number of articles mentioning a particular fact was calculated as the number of search results corresponding to a search engine query with the fact.

4 Pilot Experiment

Aside from the choice of metric, several other choices had to be made. The first choice was which search engine to use. As we had no reason to believe that a particular search engine will perform better than others, our pilot tested the metrics on the three major search engines: AltaVista (Yahoo), Bing and Google.

The second choice is what search terms to choose. Most properties can be expressed as a combination of the attribute and a value extracted from sentences such as "Alfred Nobel was born in Stockholm' (attribute: bornIn, value: Stockholm).' Choosing the value only would lead to a loss of information, because there would be no difference between properties such as ⟨bornIn : Stockholm⟩ and ⟨diedIn : Stockholm⟩ (since in both cases we would only search for Stockholm). On the other hand, attributes such as *actedIn* can be expressed by many similar expressions (e.g., starred). In such case, using both the attribute and the value might be too restrictive. As only empirical testing can show which option is better, we tested both. In the following tables, V stands for value only (e.g., "Stockholm") and AV stands for attribute and value (e.g., "born in Stockholm").

Thirdly, there is the question what to do with synonyms. While sometimes it might help to let a metric count all synonyms of a word, as people remember concepts rather than exact words, sometimes we would prefer to look for an exact phrase. This is especially the case when the value of the property is a proper name. This means that we had the option to quote the searched term to force the used search engine to look for an exact match. Again, our pilot tested both options (i.e., quoting and no quoting).

The choices described above left us with a large number of combinations. To minimise the likelihood of type II errors, we first performed a pilot experiment. The pilot used a different set of stimuli than the final evaluation experiment. Based on the pilot, we then selected the most promising combinations. The setup and the procedure used in the pilot experiment were similar to the final evaluation experiment (which is described in the following sections).

Table 1: Results of the pilot study: Spearman correlation between the heuristics and knowledge of hearers.

SE + Opt	Frequency	P(n | p)	P(p | n)	PMI
AltaVista V	0.27	0.25	0.30	0.32
Bing V	0.25	0.20	0.26	0.29
Google V	**0.47**	0.14	**0.37**	**0.51**
Google AV	**0.60**	0.23	**0.50**	**0.64**

Table 1 shows the Spearman correlations between the results of some of the metrics (unquoted

option) and people's judgement in the pilot study. The options with quoted properties proved less useful so our final evaluation used unquoted properties. The best results were achieved by using Google and expressing properties as attribute and value. Field (2009) treats values around 0.1 as indicating small effects, values around 0.3 as medium effects and values around 0.5 as large effects. This standard terminology gives our PMI and Frequency based metrics a large (positive) correlation, and our P(p | n) metric a medium (positive) correlation.

5 Evaluation

To assess the performance of a metric we performed an experiment where we presented participants with statements such as "Isaac Newton was a physicist" or "Isaac Newton was a warden of the Royal Mint" and asked them to state whether each statement was true, false or they did not know. The experiment consisted of 120 statements about 10 famous people, 7 true and 5 false statements per famous person. This set of statements was different from the pilot set. The experiment data were collected online using the Amazon Mechanical Turk restricted to US and UK workers. The participants were required to perform a short cloze test to ensure that they have a reasonable level of English. 61 human participants answered the questions about famous people. The false statements were used to ensure participants attention and were not used in the analysis. The number of correct answer to the true statements were converted to percentages. These percentages of affirmative answers (answers where participant correctly selected True) were correlated with the score of each of the metrics. This was done using the Spearman correlation measure.

Figure 1 shows the names of the famous people whose properties were used in the experiment and table 2 shows the properties of Ernest Hemingway.

- Admiral Nelson
- Alfred Nobel
- Andy Warhol
- Duke of Wellington
- Emperor Hirohito
- Ernest Hemingway
- Florence Nightingale
- Heinrich Himmler
- Louis Pasteur
- Plato

Figure 1: Famous people used in the evaluation experiment.

The best two metrics were the Frequency metric (correlation of 0.64) and the PMI-based metric (correlation of 0.66). The counts used in the calculations of scores of individual facts were estimated using Google search engine. Both correlations were significant at $p < 0.001$. More details about this experiment can be found in Kutlak, Deemter, and Mellish (2012).

Table 2: List of properties of Ernest Hemingway, corresponding condition and the percentage of affirmative answers. Rank shows how the corresponding properties ranked according to the PMI metric using Google.

Property	Condition	Percentage	Rank
Ernest Hemingway was a writer.	true	100.0	1
Ernest Hemingway was American.	true	100.0	2
Ernest Hemingway received the Nobel Prize in Literature.	true	63.6	4
Ernest Hemingway is the author of For whom the bell tolls.	true	54.5	3
Ernest Hemingway committed a suicide.	true	50.0	6
Ernest Hemingway was British.	false	27.3	-
Ernest Hemingway was born in Oak Park.	true	25.0	5
Ernest Hemingway received the Italian Silver Medal of Bravery.	true	20.0	7
Ernest Hemingway is the author of A tale of two cities.	false	13.3	-
Ernest Hemingway invented dynamite.	false	0.0	-
Ernest Hemingway died in a plane crash.	false	0.0	-
Ernest Hemingway was born in Paris.	false	0.0	-

6 Conclusion

We hypothesised that corpora can be used to estimate what facts are in common ground of a particular community. We evaluated several collocation measures to test our hypothesis and the results of our experiment suggest that the Frequency and the PMI metrics perform well in this task and can serve as a heuristic for estimating what facts are likely to be know by the community of English speakers. Our next step is to evaluate the heuristic in the actual task of selecting the content of descriptions of famous people.

7 Acknowledgements

We would like to thank the members of the University of Aberdeen Natural Language Generation group and the reviewers for their valuable comments. This research is sponsored by the Scottish Informatics and Computer Science Alliance (SICSA).

References

Clark, H. H. (1996). *Using language*. New York: Cambridge University Press.

Fano, R. M. (1961). *Transmission of information: A statistical theory of communications*. New York: Wiley.

Field, A. (2009). *Discovering statistics using spss*. SAGE publications Ltd.

Grice, P. (1975). Logic and conversation. *Syntax and Semantics*, *3*, 43–58.

Jucks, R., Becker, B.-M., & Bromme, R. (2008). Lexical entrainment in written discourse: Is experts' word use adapted to the addressee? *Discourse Processes*, *45*(6), 497-518.

Kutlak, R., Deemter, K. van, & Mellish, C. (2012). Corpus-based metrics for assessing communal common ground. *To appear in: Proceedings of the 34th Annual Meeting of the Cognitive Science Society*.

Nickerson, R. S., Baddeley, A., & Freeman, B. (1987). Are people's estimates of what other people know influenced by what they themselves know? *Acta Psychologica*, *64*(3), 245 - 259.