Semantic similarity and the generation of referring expressions: A first report

Albert Gatt & Kees van Deemter
Department of Computing Science, University of Aberdeen

{agatt | kvdeemter}@csd.abdn.ac.uk

1 Motivation

The past decade[1], has witnessed renewed interest in the Generation of Referring Expressions (GRE) [23, 24, 8, 9, 10, 12, 22]. Broadening the scope beyond earlier work [3, 4, 5], recent proposals involve algorithms that refer to sets as well as individuals, using operations such as set union ('the cat and the dogs') and complementation ('the dog that is not black'). As a consequence, it has become more difficult for a generator to choose among alternative expressions that may be coextensive. This paper is part of a concerted effort to shed some empirical light on the question of expressive choice. The focus is on reference to sets, where a referring expression is built by unifying two or more singletons. Starting with descriptions of the form 'the N_1 and (the) N_2', we investigate whether the semantic similarity of N_1 and N_2 is relevant in determining the acceptability of the generated NP.

Suppose that, in a given domain, an entity e_1 can be referred to as either 'the postgraduate' or 'the psychologist'; similarly, e_2 can be referred to as either 'the undergraduate' or 'the man on the first floor'. Various alternatives exist for an expression referring to $\{e_1, e_2\}$, e.g.: (i) 'the postgraduate and the man on the first floor', (ii) 'the postgraduate and the undergraduate', (iii) 'the psychologist and the undergraduate'. Here, (ii) is arguably better than (i) or (iii). Intuitively, this is because the conjuncts in (ii) are more semantically similar or 'related'. Moreover, expression (iii) violates the Gricean maxims. The choice of two equally specific [2] but semantically unrelated descriptors, 'psychologist' for e_1 versus 'undergraduate' for e_2, might give rise to (false) implicatures, such as that the two entities have nothing in common, thus violating the Gricean Cooperative Principle, and resulting in a description which is less coherent than it might be. Suppose further that e_1 and e_2, as well as a third entity e_3 referred to as 'the book', were introduced in a discourse. Subsequent reference to a pair of these entities might be made via a coordinate construction, or some other structure. Considerations of semantic similarity may guide the choice between alternatives; in particular, referring to the set $\{e_1, e_2\}$ using an NP conjunction is more felicitous than a similar reference to $\{e_1, e_3\}$ ('the psychologist and the book'). In the latter case, it may be more felicitous to refer to these two entities using different phrases. A third consideration has to do with a user's comprehension of a generated text. If a description gave rise to false implicatures, or simply sounded odd as a result of an infelicitous choice of descriptors, the quality of the text and its comprehensibility would be reduced. We next describe a correlational study which investigated the relationship of semantic similarity and perceived acceptability of conjoined NPs. Our study is closely related in spirit to [13], which also

[1]This work is part of the TUNA project, which is funded by the UK's Engineering and Physical Sciences Research Council (EPSRC) under grant number GR/S13330/01.
See `http://www.itri.brighton.ac.uk/projects/tuna/index.html`

evinces a concern with semantic plausibility and its implications for NLG, albeit in a different domain.

2 The experiment

The experiment was conducted over the World Wide Web using Magnitude Estimation (ME) to elicit acceptability judgments [1]. In ME, participants are first asked to rate a first item, called the *modulus*, on a numeric scale of their own choice; all successive items are compared to the modulus for acceptability. A total of 63 self-reported native or fluent speakers completed the experiment in exchange for participation in a prize draw. Three were later excluded from analysis due to a failure to follow instructions. Instructions emphasized acceptability as a function of the *likelihood of usage of a phrase in some situation*. After rating the modulus phrases, trials consisted of a page displaying the original modulus and rating, plus a new phrase. Two within-subjects factors were manipulated: *frequency* and *phrase-type*. Nouns extracted from the BNC with frequency counts were lemmatised using the Sussex Morphological Analyser [18], and split into four frequency bands ranging from *High* ($f > 500$ per million) to *Low* ($f < 100$ pm). From each frequency band, 16 nouns were randomly selected and paired. The second factor, *phrase-type*, manipulated whether a phrase had the form *the N and N* (1-det condition) or *the N and the N* (2-det condition). Four pairs of nouns were assigned to each *phrase-type* condition within each frequency band, yielding a 2×4 within-subjects design. To ensure comparability of ratings across subjects and normalisation of data, analysis took into account the mean preference for each trial, where preference is calculated as $log(t/m)$ for a trial value t and modulus value m.

2.1 Semantic similarity measures

Four measures of semantic similarity were considered, the first three based on WordNet and frequency information from the BNC. These were generated using methods from the WordNet::Similarity Perl module [19]. Here we focuse exclusively on similarity between the primary WordNet sense of two nouns/concepts, denoted a, b.

1. *Resnik* [20]: Augments the taxonomy with a monotonically increasing function $p : c \rightarrow [0, 1]$, where $p(c)$ is the probability of encountering an instance of concept c which subsumes a and b. $p(c)$ increases the further up the taxonomy c is, with a corresponding decrease in information content.

2. *Lin* [16]: An information-theoretic measure calculated as a function of (i) the information content of the least common subsumer of a and b; (ii) the information content of a description of a and b.

3. *Minimum Path* [19]: In this measure, the similarity of a and b is the inverse of the shortest path length between the two concepts.

4. *WASPS* [11]: An adaptation of [16] to the calculation of similarity of words, based on their occurrence in dependency triples in corpora. The WASPS theasaurus used here was constructed by Kilgariff and Tugwell from the British National Corpus.

2.2 Results and discussion

In a by-items ANOVA, neither *frequency* ($F < 1$; $p > .5$) nor *phrase-type* ($F < 1$; $p > .7$) was found to be significant. Though the lack of a frequency effect is surprising, given that it has been frequently attested (e.g. [13]), it can be explained by the fact that participants were rating acceptability of phrases which consisted of pairs that were matched for frequency. The

lack of a frequency effect reduces the possibility that acceptability ratings of these NPs could be modulated by the frequency or familiarity of the nouns involved.
Correlations were generated between the mean preference on each trial and the four similarity measures, as well as a random number in $(0, 1)$. The least correlated measure was Minimum Path (*Pearson's* $r = .480,\ p = .05$). The highest correlations were obtained for the WASPS ($r = .576,\ p < .01$) and Resnik measures ($r = .538,\ p < .01$), followed by the Lin measure ($r = .444,\ p > .01$). Crucially, the random measure did not correlate significantly ($r = .246$, p negligible). The two most significantly correlated measures are both corpus-based. For the Resnik measure, $p(c)$ is calculated on the basis of frequency information obtained from a corpus; WASPS is a corpus-based measure which implicitly accounts for distinctions of word meaning as evinced by the occurrence of words in similar grammatical environments. Although the corpus-based measures do not exhaust the possibilities, given the relatively high correlation of the taxonomy-based Lin measure, speakers' intuitions seem to be influenced by a similarity metric which is, at least in part, determined by distributional and quantitative information. Assuming that a corpus such as the BNC is a representative snapshot of language use at a particular time, this supports the idea that linguistic acceptability is strongly influenced by the degree of exposure to certain constructions. This is also good news from a practical point of view, since it suggests that reliance on hand-crafted taxonomies such as WordNet is not always necessary.

3 Conclusions and future work

Our findings suggest that semantic similarity is significantly correlated to the perceived acceptability of NP conjunctions. From the generation point of view, we have suggested that this finding is relevant both for choice of descriptors during content determination and for subsequent aggregation/realisation. An interesting perspective on these findings can be found in recent work in psycholinguistics. Empirically-grounded models of lexical access in speech production [7, 21] have converged on the view that the mental lexicon involves an associative memory, where concepts which are semantically related are adjacent (cf. [14] for a review). Evidence for this comes, for example, from the *semantic interference* effect, where the production of a target word is inhibited by the presence of a semantically similar distractor [25, 6]. This is often explained with reference to spreading activation: Semantic similarity results in concepts (and hence, lemmas) for both target and distractor being activated. The work reported here raises two interesting questions. The first is related to the definition of semantic similarity which should inform the structure of such models. We have suggested that this definition should be corpus-based and distributional rather than purely taxonomic (cf. the weak correlation for the Minimum Path measure). The second question is more directly related to language production/generation. Given that speakers' intuitions favour semantically similar descriptors in complex NPs, will this also be reflected in production scenarios, in which speakers need to refer to objects and have a choice between a range of descriptors? Our findings would lead to the prediction that it would. This might be explained in terms of *auto-priming*, in which the selection of an initial descriptor (say, for entity e_1) makes it easier to retrieve a second descriptor which is closely associated. These are open questions, given the relative lack of work on the production of complex NPs such as conjunctions (but see [15, 17]), and the preliminary nature of the study reported here. We are using these results to inform more controlled experiments investigating the effects of similarity on descriptor choice in a reference task, with a view to implementation of these heuristics as part of a GRE algorithm.

References

[1] Bard, E.G., Robertson, D., and Sorace, A. (1996). Magnitude estimation of linguistic acceptability. Language, 72(1): 32-68.

[2] Cruse, D.A. (1977). The pragmatics of lexical specificity. Journal of Linguistics, 13: 153-164.

[3] Dale, R. (1989). Cooking up referring expressions. Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, ACL-1989.

[4] Dale, R., and Haddock, N. (1991). Generating referring expressions containing relations. Proceedings of the 5th Conference of the European Chapter of the ACL, EACL-1991.

[5] Dale, R., and Reiter, E. (1995). Computational interpretation of the Gricean maxims in the generation of referring expressions. Cognitive Science, 19(2): 233-263.

[6] Damian, M.F., Vigliocco, G., and Levelt, W.J.M. (to appear). Effects of semantic context in the naming of pictures and words. Cognition.

[7] Dell, G.S. (1986). A spreading-activation theory of retrieval in sentence production. Psychological Review, 93: 283-321.

[8] Gardent, C. (2002). Generating minimal definite descriptions. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-2002.

[9] Horacek, H. (2003). A best-first search algorithm for generating referring expressions. Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, EACL-2003.

[10] Horacek, H. (2004). On referring to sets of objects naturally. Proceedings of the 3rd International Conference on Natural Language Generation, INLG-2004.

[11] Kilgarriff, A., and Tugwell, D. (2001). WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. Proceedings of the Collocations Workshop in Association with ACL-2001.

[12] Krahmer, E., van Erk, S., and Verleg, A. (2003). Graph-based generation of referring expressions. Computational Linguistics, 29(1): 53-72.

[13] Lapata, M., McDonald, S., and Keller, F. (1999). Determinants of adjective-noun Plausibility. Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics, EACL-1999.

[14] Levelt, W.J. (1998). Models of word production. Trends in Cognitive Science, 3(6): 223-232.

[15] Levelt, W.J., and Maassen, B. (1981). Lexical search and order of mention in sentence production. In W. Klein and W.J.M. Levelt (Eds.), Crossing the boundaries in linguistics (pp.221-252). Dordrecht: Reidel.

[16] Lin, D. (1998). An information-theoretic definition of similarity. Proceedings of the International Conference on Machine Learning.

[17] Meyer, A.D. (1996). Lexical access in phrase and sentence production: Results from picture-word interference experiments. Journal of Memory and Language, 35: 177-196.

[18] Minnen, G., Carroll, J., and Pearce, D. (2001). Applied morphological processing of English. Natural Language Engineering, 7(3): 207-223.

[19] Pederson, T., Patwardhan, S., and Michelizzi, J. (2004). WordNet::Similarity — Measuring the relatedness of concepts. Proceedings of the Nineteenth National Conference on Artificial Intelligence, AAAI-2004.

[20] Resnik, S. (1995). Using information content to evaluate semantic similarity in a taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI-1995.

[21] Roelofs, A. (2000). WEAVER++ and other computational models of lemma retrieval and word-form encoding. In L. Wheeldon (Ed.), Aspects of Language Production. Sussex, UK: Psychology Press.

[22] Siddharthan, A., and Copestake, A. (2004). Generating Referring Expressions in Open Domains. Proceedings of the 42th Meeting of the Association for Computational Linguistics Annual Conference, ACL-2004.

[23] Stone, M. (2000). On identifying sets. Proceedings of the 1st International Conference on Natural Language Generation, INLG-2000.

[24] van Deemter, K. (2002). Generating Referring Expressions: Boolean Extensions of the Incremental Algorithm. Computational Linguistics, 28(1): 37-52.

[25] Vigliocco, G., Lauer M., Damian M.F., and Levelt W.J. (2002). Semantic and syntactic forces in noun phrase production. Journal of Experimental Psychology: Learning, Memory and Cognition, 28(1): 46-58.