TECHNICAL NOTE

Evaluation of ENIGMA's Lexical Choice Component

David Hardcastle

http://www.davidhardcastle.net

July 2007

Evaluation of ENIGMA's Lexical Choice Component

David Hardcastle
Open University, Milton Keynes
Birkbeck College, London

Abstract

ENIGMA is a Natural Language Generation (NLG) system that generates cryptic crossword clues. A key component of the system determines lexical choice decisions, in other words it decides which actual word to use to communicate keyword concepts within the clue. For example if a word is to be used as an anagram in the wordplay puzzle communicated by the clue, then ENIGMA must choose from a list of anagram keywords one that could be meaningfully combined with the anagram in English prose. The experiment described in this technical note was designed to test whether the choices made by this component could pass for fragments of natural language.

Introduction

ENIGMA is an NLG system that generates cryptic crossword clues and forms the background to my PhD thesis. The focus of this technical note is the evaluation of the lexical choice component of ENIGMA, the part of the system that determines which words to use in a given context. ENIGMA explores a wide range of options for constructing the elements of a crossword clue and expressing it in English, and then applies a range of constraints to limit those choices (see Hardcastle, 2007). These constraints apply to clue formulation, organisation and realization and include domain-specific information relating to crossword conventions, and also generic information relating to plausibility in natural language.

One such constraint applies to the relationships that exist between the words in the clue: given a particular syntactic relationship there is a semantic constraint on which words can fill slots in the syntactic frame. For example, if ENIGMA intends to construct a noun phrase *x pear*, where *x* is treated as an adjective and *pear* as a noun, then there are semantic constraints on which adjectives *x* can be. In this example adjectives such as *rotten*, *sweet*, *green* or *soft* would all be acceptable, whereas adjectives such as *liberated*, *clever* or *noisy* would not. There are also many adjectives that are not obvious collocates for *pear*, but could nonetheless be applied to a pear in some possible situation, such as *large*, *unusual*, *squashed* or *fresh*, and which probably relate broadly to the features that pears are thought to possess.

To determine which adjectives are appropriate for a noun such as *pear* ENIGMA uses a 'collocational semantic lexicon' extracted from the British National Corpus and generalized using WordNet. The details of the extraction process are described in (Hardcastle, 2007b). An example entry from the lexicon is given below.

naval
generalizations: training aviator gathering army_unit military_action activity unit team vessel war military_officer commissioned_military_officer warship ship military_unit body
single entries: engineer midshipman cadet yard uniform helicopter frogman aviator supremacy presence tradition academy tactic victory expert strength chronicle port fighter superiority auxiliary shipyard capability historian rating policy unit fleet people expenditure code activity defeat hospital store surgeon observatory
collocations: naval base, naval officer, naval dockyard, naval architect, naval gun, naval reserve, naval vessel

Figure 1. A sample entry from the collocational semantic lexicon. The entry is for the adjective *naval*, it asserts that any hyponym of the generalizations such as *aviator* or *military_action* can be said to be naval, as could any of the words listed as single entries. The lexicon also proposes collocations such as naval base.

Experiment

Evaluating the semantic collocational lexicon was difficult, since the associations asserted by the lexicon relate not just to linguistic but also to real-world information. The fact that *lamb* can be *grilled*, for example, tells us something about the real world interaction of the noun *lamb* and the act of *grilling* something, and there are no resources that would support an automated quantitative evaluation of the lexicon. I therefore constructed a task-based evaluation in which ENIGMA performed a standard crossword compilation sub-task, assigning an adjective that means "is an anagram" in crosswords to a noun that is an anagram of some other word, and turned to a sample of native and non-native speakers of English to act as adjudicators. Note that no domain expertise is required; the evaluators are not being asked to judge whether the adjectives can mean "is an anagram" in crosswords, only whether the adjective and the noun combine to form a reasonable fragment of surface text.

Many of the subjects reported that they found the task difficult; in the absence of context it was difficult to make judgements about examples that were not prototypical, especially when there was no clear winner and they felt that they were being asked to make quite fine-grained distinctions. This underlines the how hard this task is for a computer program since there are many grey areas and matters of judgement.

I chose to limit the task to one of the three dependency relations that I used as the basis for the lexicon in order to keep the task clear and simple to explain. Given the feedback from many of the subjects I think that this was the right choice.

Subjects

Thirty-two subjects took part in the experiment, of whom six are non-native speakers of English. Most have an interest in crosswords, NLP, dictionaries or publishing, and so, like ENIGMA's target audience (crossword solvers), most of the subjects could reasonably be said to have an interest in words and language.

Methodology

To assess the choices made by ENIGMA I presented each subject with a forced choice in which they had to choose from a list of three adjectives presented alongside a noun to form the adjective-noun combination they thought most likely to occur in spoken English. This design is similar to the fill-in-the-blanks test used to assess linguistic competence except that the collocation type is narrowed to a single relation and the context has been removed. In standard fill-in-the-blanks tests the subjects use the collocation and the context to choose the most appropriate word. Although ENIGMA will also give consideration to consistency of context across the clue, that component of the system is not under test here, and so the experiment focuses solely on the collocation. As above, it seems that this is a much harder task, even for native speakers, reinforcing the notion that meaning is not just located within words and their immediate dependencies.

The test comprises sixty nouns chosen at random from a list of whole word anagram nouns. Each noun is then accompanied by three adjectives. One is chosen at random from the anagram keyword adjectives that ENIGMA selects for that noun, the other two are chosen at random from the remaining anagram keyword adjectives. In thirty cases the choice of adjective from ENIGMA's list was biased by adding adjectives where the noun had been evidenced directly rather than inferred through generalization into the drawing pot twice, making them more likely to be selected. This replicates the promotion of evidenced adjectives over inferred adjectives that will occur at runtime. The biased list returned better results than the unbiased list, as one might expect. In many cases there was at least one option that was flagged as a possible collocation in the list of alternatives available to ENIGMA. I could have set the system up so that the best adjective is always chosen, ranking collocations over directly evidenced terms, and ranking these over generalizations. However, such an approach would restrict lexical variety and make the clues become formulaic and dull. The point here is not simply to come up with an adjective that can mean anagram that is a good fit to each noun, but for the system to recognise that a wide range of adjectives can fit with a noun, some better than others, offering variety in lexical choice, and also opening up more possibilities for interaction with other components of the clue.

It is of course possible that the adjectives chosen for control might themselves be a good fit to each noun. However, since they must be selected from the remaining pool of adjectives not selected by ENIGMA the system's performance has some impact on what control adjectives are chosen: if the coverage of the lexicon is sufficiently high there should be no appropriate adjectives left to choose. In practice this was not always the case, leaving an element of subjectivity open in the interpretation of negative results.

The list of nouns was presented to the subjects in a spreadsheet with a drop-down to the left of each entry. The drop-down contained the ENIGMA adjective and the two control adjectives in random order. Subjects were asked to choose the adjective-noun combination in each case that they thought they would be most likely to encounter in spoken English. I hoped that this would encourage them to think about naturalness rather than consider the options in narrow terms. The assumption behind the experiment is that this criterion of naturalness is the same criterion that will be applied when someone reads a clue and assesses how it reads as a fragment of natural language. Other features of the clue are not under test here, so for example it doesn't matter whether the anagram is a good or bad one, indeed in some cases (due to an oversight in the code that generated and filtered the data) some 'anagrams' in the test set are actually just alternate spellings of other dictionary words.

Results

The choices made by the subjects were treated as a binomial distribution in which two responses are possible: the subject may choose the same adjective that ENIGMA chose ($p = 0.33$), or they may not ($q = 1-p = 0.67$). The null hypothesis is that ENIGMA's selection has not biased the choice in any way, and so the selections will be made by the subjects on a purely random basis. I calculated the P-value for each entry as the probability under the binomial distribution of getting more than the number of matches observed if choices were made randomly, and used a confidence threshold of $p<0.01$. The null hypothesis was rejected in 51 (85%) of the 60 entries, with a marked difference between the performance of the biased set (97%) and the unbiased set (73%). This subset data suggests that the greatest accuracy would be achieved by making the algorithm more selective, but this would come at a cost to lexical variety, and so in practice a trade-off exists in the application parameters between these two requirements.

The results for all sixty entries are shown in Tables 1 and 2. Each row contains the target noun and the three adjectives presented to the user, with ENIGMA's choice in the left hand column. The percentages represent the percentage of subjects who picked each adjective, and the highlighted adjective in each row represents the most popular choice for that entry. Significant positive results are those where more than around 52% of the subjects chose the same adjective as ENIGMA.

Noun	ENIGMA		Alternative-1		Alternative-2	
bell	broken	100%	fabricated	0%	wandering	0%
spice	mixed	100%	ordered	0%	modified	0%
theorem	bizarre	100%	mutilated	0%	minced	0%
anvil	broken	96%	corrected	3%	blended	0%
sausage	boiled	96%	curious	3%	zany	0%
slip	careless	96%	fiddly	3%	spun	0%
orphan	awkward	93%	ordered	6%	assorted	0%
stable	broken	93%	rectified	6%	fiddled	0%
teenager	awkward	93%	varied	3%	untrue	3%
brand	odd	90%	warped	10%	minced	0%
itch	vague	90%	twisty	10%	fancy	0%
slice	fresh	90%	haphazard	10%	organised	0%
coating	fresh	86%	ruptured	13%	wandering	0%
reward	dreadful	86%	processed	13%	malformed	0%

sanction	curious	86%	translated	10%	runny	3%
summerhouse	broken	86%	regulated	10%	flurried	3%
microbiology	essential	83%	extraordinary	16%	poached	0%
shriek	confused	83%	rioting	16%	processed	0%
bombing	criminal	76%	suspect	23%	assembled	0%
realization	fresh	76%	naughty	16%	rectified	6%
enumeration	curious	73%	dreadful	20%	rum	6%
peacetime	exciting	73%	distressed	20%	beaten	6%
brief	original	70%	reviewed	16%	running	13%
wrong	extraordinary	70%	essential	16%	doctored	13%
handshake	curious	66%	extraordinary	30%	untidy	3%
misbehaviour	criminal	60%	revolting	26%	wayward	13%
manse	broken	56%	mutilated	26%	derived	16%
bloodshed	free	53%	corrupted	33%	naughty	13%
still	original	53%	doctored	36%	yawing	10%
quiet	awkward	50%	disturbed	40%	spoilt	10%

Table 1. Results for the thirty biased entries.

Noun	ENIGMA		Alternative-1		Alternative-2	
bore	dreadful	96%	malformed	3%	adapted	0%
follow-on	possible	96%	rum	3%	broiled	0%
lid	loose	96%	reformed	3%	running	0%
praise	vague	96%	formed	3%	broiled	0%
scowl	odd	96%	reformed	3%	brewed	0%
bass	fresh	93%	doctored	3%	developed	3%
shit	bad	93%	abnormal	6%	organised	0%
dearth	possible	90%	invented	6%	wrecked	3%
fart	swirling	90%	varied	6%	forged	3%
trail	fresh	90%	cracked	6%	ill-formed	3%
footprint	abnormal	86%	bad	13%	used	0%
stranglehold	possible	86%	converted	6%	fluid	6%
trace	curious	86%	compounded	13%	broadcast	0%
allure	extraordinary	83%	made-up	16%	converted	0%
lunatic	potential	80%	converted	16%	original	3%
default	complicated	76%	assorted	13%	ruined	10%
hen	minced	76%	jumbled	23%	reviewed	0%
olive	abnormal	76%	random	23%	disrupted	0%
saint	poor	76%	suspect	23%	arranged	0%
moral	free	63%	hashed	26%	repaired	10%
pathos	original	63%	riotous	23%	volatile	13%
hare	chopped	60%	tricky	33%	warped	6%
pulse	cooked	40%	dizzy	40%	fantastic	20%
sorting	free	40%	perverse	56%	amiss	3%
honey	ground	23%	dressed	56%	damaged	20%
alto	extraordinary	20%	shaky	80%	entangled	0%
knockout	free	13%	staggering	86%	mutilated	0%
mixer	scattered	13%	novel	73%	arranged	13%
oath	incorrect	13%	complicated	86%	cooked	0%
goodwill	ill	3%	suspect	73%	broadcast	23%

Table 2. Results for the thirty unbiased entries.

Baseline Comparison

To see how the generalized data extracted via dependencies improved on cooccurrence data I used the Russian Doll algorithm, a word association measure based on a statistical analysis of cooccurrence data from the British National Corpus described in (Hardcastle, 2005), to generate a baseline for comparison. The association score for each adjective when combined with the test noun is shown in Table 3 below - pairings scoring above the threshold of 0.40 are highlighted. The Russian Doll algorithm selected at least one of the three possibilities in 14 of the 60 sample entries, and in 11 of these 14 cases the adjective favoured by the participant group scored over the threshold. In four cases there were two adjectives that passed the threshold, and in three cases the chosen adjective did not match the choice made by the group. This shows that it is possible for a system based on cooccurrence data to replicate the performance of the collocational semantic lexicon some of the time, but because the data is not generalized it lags far behind in coverage.

Noun	Adjective-1		Adjective-2		Adjective-3	
bass	doctored	0.00	developed	0.00	fresh	0.04
microbiology	poached	0.00	extraordinary	0.00	essential	0.27
spice	ordered	0.05	mixed	1.00	modified	0.00
brand	odd	0.08	warped	0.11	minced	0.00
slice	haphazard	0.01	fresh	0.72	organised	0.00
wrong	doctored	0.00	extraordinary	0.22	essential	0.30
bell	broken	0.43	fabricated	0.00	wandering	0.07
alto	extraordinary	0.12	entangled	0.00	shaky	0.72
enumeration	rum	0.00	dreadful	0.00	curious	0.00
bloodshed	corrupted	0.38	free	0.14	naughty	0.00
misbehaviour	criminal	0.45	wayward	0.98	revolting	0.25
summerhouse	broken	0.38	regulated	0.00	flurried	0.00
sanction	runny	0.00	curious	0.01	translated	0.00
bombing	criminal	0.24	suspect	0.11	assembled	0.19
shit	abnormal	0.00	organised	0.00	bad	0.31
stranglehold	converted	0.00	possible	0.05	fluid	0.00
saint	poor	0.21	suspect	0.07	arranged	0.17
peacetime	distressed	0.05	exciting	0.00	beaten	0.00
still	original	0.29	yawing	0.00	doctored	0.00
trail	fresh	0.27	ill-formed	0.00	cracked	0.17
pulse	fantastic	0.18	dizzy	0.46	cooked	0.47
slip	spun	0.08	careless	0.62	fiddly	0.00
handshake	extraordinary	0.35	untidy	0.06	curious	0.27
default	assorted	0.00	ruined	0.00	complicated	0.29
realization	rectified	0.00	naughty	0.02	fresh	0.13
anvil	corrected	0.00	blended	0.00	broken	0.20
pathos	volatile	0.16	original	0.10	riotous	0.40
teenager	varied	0.16	untrue	0.28	awkward	0.55
oath	cooked	0.01	incorrect	0.03	complicated	0.04
quiet	disturbed	0.22	awkward	0.20	spoilt	0.00
stable	broken	0.29	rectified	0.00	fiddled	0.00
theorem	mutilated	0.00	bizarre	0.00	minced	0.00
coating	ruptured	0.00	fresh	0.18	wandering	0.00
follow-on	rum	0.00	possible	0.08	broiled	0.00
fart	swirling	0.00	forged	0.00	varied	0.00
lid	running	0.13	loose	0.52	reformed	0.00
manse	mutilated	0.00	broken	0.39	derived	0.00

hen	reviewed	0.00	jumbled	0.00	minced	0.05
allure	made-up	0.00	extraordinary	0.24	converted	0.00
footprint	abnormal	0.00	bad	0.05	used	0.04
praise	formed	0.00	broiled	0.00	vague	0.20
goodwill	suspect	0.00	ill	0.25	broadcast	0.00
knockout	staggering	0.00	mutilated	0.00	free	0.09
dearth	possible	0.23	invented	0.00	wrecked	0.16
scowl	brewed	0.00	odd	0.15	reformed	0.00
olive	abnormal	0.00	random	0.00	disrupted	0.23
shriek	processed	0.00	rioting	0.00	confused	0.23
hare	tricky	0.08	chopped	0.00	warped	0.49
trace	compounded	0.00	curious	0.21	broadcast	0.00
honey	damaged	0.07	dressed	0.69	ground	1.00
bore	malformed	0.00	adapted	0.00	dreadful	1.00
itch	vague	0.18	twisty	0.00	fancy	0.10
mixer	novel	0.00	scattered	0.27	arranged	0.51
orphan	assorted	0.06	awkward	0.13	ordered	0.00
moral	repaired	0.00	hashed	0.00	free	0.17
brief	original	0.46	reviewed	0.00	running	0.66
lunatic	potential	0.05	converted	0.00	original	0.32
sorting	perverse	0.00	amiss	0.00	free	0.12
sausage	boiled	0.87	curious	0.11	zany	0.00
reward	dreadful	0.09	processed	0.02	malformed	0.00

Table 3. Scores for each adjective against the noun using the Russian Doll algorithm described in (Hardcastle, 2005). The threshold is 0.40.

Discussion

It is interesting to note that in the case of the nine negative results, in which the null hypothesis was not rejected, there are only two entries where it seems that the subjects were unable to agree on the most apposite adjective. In the remaining seven cases there was significant agreement amongst the subjects, but this agreement did not chime with the choice made by the system.

This means that there are two distinct findings that can be drawn from the set of twelve negative results. There is one group in which none of the three adjectives presented to the subjects emerges as the clear favourite, and the distribution appears to be close to random. In these cases, listed in Table 4 below, it is clear that the adjective selected from the lexicon is simply not an apposite modifier for the noun received as input. However, there is another group in which there was a clear favourite, but the favourite is not the adjective selected by the system. The interpretation of *these* negative results is rather subjective. It could be that ENIGMA chose a good modifier, but the control set included an even better one, or it could be that the control set included a passable modifier and the adjective chosen by ENIGMA was not up to scratch. These six entries are shown in Table 5.

Noun	ENIGMA		Alternative-1		Alternative-2	
quiet	awkward	50%	disturbed	40%	spoilt	10%
pulse	cooked	40%	dizzy	40%	fantastic	20%

Table 4. Negative results in which there was no significant favourite.

Noun	ENIGMA		Alternative-1		Alternative-2	
sorting	free	40%	perverse	56%	amiss	3%
honey	ground	23%	dressed	56%	damaged	20%
alto	extraordinary	20%	shaky	80%	entangled	0%
knockout	free	13%	staggering	86%	mutilated	0%
mixer	scattered	13%	novel	73%	arranged	13%
oath	incorrect	13%	complicated	86%	cooked	0%
goodwill	ill	3%	suspect	73%	broadcast	23%

Table 5. Negative results in which there was a significant favourite, but it differed from the choice made by the system.

This second set of negative results cannot be explained away; under the terms of the evaluation they are negatives and so the system performance remains at 85%. However, there are some interesting points to make in some cases, that suggest that circumstantial factors rather than a failure on the part of the system may have caused the subjects to agree on a different adjective than the choice made by ENIGMA.

Polysemy in the noun

The entry for *pulse* is interesting. Most subjects selected *cooked*, presumably on a reading of *pulse* meaning an edible seed, whereas many also chose *dizzy*, presumably relating it to a reading of *pulse* meaning the rhythm of the arteries. The same may be true for *dressed honey* where the option provided by ENIGMA is based on the sort of honey that is made by bees, although of course it cannot in fact be *ground*, whereas the subjects may have thought of *honey* as an attractive woman, although this would also depend on how they read the polysemy in *dressed* so we cannot be certain.

Word association

Dizzy pulse is also interesting as, it seems to me at least, that a pulse itself cannot be dizzy, although both dizziness and one's pulse can be closely connected. The same could perhaps be said of *staggering knockout*. Perhaps when faced with a forced choice the subjects at times will fall back not on collocational information based on dependencies, but on thematic associations of the sort derived by a bag of words analysis.

Real world knowledge

Several of the choices made by ENIGMA had a culinary theme: *boiled sausage*, *minced hen*, *chopped hare* and *ground honey*. The first three are certainly not prototypical treatments of the ingredient in each case, but they were readily agreed on by the subjects. However, the fourth was largely rejected, perhaps because it had crossed a line into an assertion that is not physically possible, one cannot grind a liquid substance, rather than one that is merely unusual or faintly implausible.

Apposite Alternatives

In a couple of cases an alternative was simply very apposite. *Shaky alto* was much preferred to *extraordinary alto*, and *complicated oath* was preferred to *incorrect oath*, although in each case ENIGMA's selection appears to me, at least, to be valid. As discussed above this demonstrates limits to ENIGMA's coverage; ideally *shaky* would not have been available for random selection as ENIGMA would have marked it as a possible fit to *alto*.

Overall then the results are very encouraging. Although the process of aligning adjectives to nouns without direct precedent involves trying to map elements of semantics that include some implicit knowledge about the real world the system got it right 85% of the time. Looking at the negative results in more detail also suggests that where ENIGMA was not successful in matching the subjects' choice there may be other factors at work since in most cases there was agreement among the subjects, leaving a subjective judgement open as to the reasonableness of the selection made by the system in some of the negative results.

References

Hardcastle, D. (2007). *Generalizing Syntactic Collocates for Creative Language Generation.* In Proceedings of Using Corpora in NLG and Machine Translation Workshop at MT Summit XI, Copenhagen, Denmark.

Hardcastle, D. (2007). *Cryptic Crossword Clues: Generating Text with a Hidden Meaning.* In Proceedings of 11th European Workshop on Natural Language Generation, Schloss Dagstuhl, Germany.

Hardcastle, D. (2005). *An examination of word association scoring using distributional analysis in the British National Corpus: what is an interesting score and what is a useful system?* In Proceedings of Corpus Linguistics 2005, Birmingham, UK.