1 John Barnden's Comments and Respones

Metaphor and related things (mainly section 6.1)

JB-1: Make it clear that in the current state of play the system only responds to pre-existing similarity, rather than being capable of doing things like adding new features to the target based on features of the source (where in many cases the new features can contravene existing knowledge of the target). As far as I understand it, with "surgeons are butchers", the vector for surgeons is not affected by the metaphor, and also no meaning vector is constructed - cf. the meaning vector that Utsumi's vector based approach to metaphor computes on the basis of the source and target vectors.

> *Yes, my approach involves the discovery of a subspace which is in some sense salient to both source and target, and so it would be wrong to claim that I'm identifying the properties transferred from target to source, or generating a conceptually revised version of the target vector. I've added a paragraph clarifying this.*

Optionally you could add something to the future work part of the thesis about prospects the system has for the above sort of thing or interpretation in other respects.

> *I've added a couple sentences about the potential for using my methodology for both interpretation and generation of metaphor.*

JB-2: Make it clearer that as it stands the method is not sensitive to information provided by small local contexts that might affect what it does with something (e.g. a metaphor) placed within that context. Or, of course, if it is sensitive to it in some way, make the matter clear.

> *I've included a sentence regarding the insensitivity to highly specific informational transfer along with the paragraph added to clarify the lack of interpretability of conceptual transfer added in response to the previous comment.*

Coercion (section 6.2)

JB-3 : p.137f: You seem to be just using the context words as subspace selectors, ignoring the targeted pair itself, and if I remember correctly you confirmed this in the viva. But why not compare performance between

(a) original method, using the word pair but not the context words, with

(b) that method plus context words?

This should surely reveal the effect of context in a better way? I may be wrong, but anyway add something about whether it would be a good approach or not.

> *I've added a short paragraph about this useful suggestion for further research (including a very brief reflection on some potential problems and a potential solution).*

Metaphor/Coercion Ramifications (section 6.3)

JB-4: p.141 bottom: The claim that there is a strong correlation between figurativeness and the need to process context during interpretation needs additional justification (or weakening of the claim if you can't comply).

> *I have both tempered the claim and cited psycholinguistic research indicating increased processing time for increasingly unfamiliar metaphors.*

JB-5: p.143 last para: Do you mean you can compute new meaning by moving around within the spaces? If so, this could be something to mention in emendments responding to my comment (1) in the Metaphor segment above.

> *In hindsight I think I was a bit vague here: meaning is not so much about moving around in a space as the discovery of opportunities for communication in the geometry of spaces induced by a certain context. I've added a sentence to clarify this.*

Analogy (Ch7)

JB-6: Re an analogy experiment, around p.151: Add commentary on why seemingly irrelevant dimensions/terms such as ”hearing” and ”accidentally” come into play, and why relevant dimensions don't appear, such as perhaps ”sit” in the analogy involving sofas. What *is* the intuitive significance of the most prominent dimensions that your analysis uncovers? I didn't understand your answer in the viva that it doesn't matter what the dimensions are. So considerable extra clarification is needed on these matters.

> *Yes, this is a good and important question. The answer is that there isn't necessarily any immediately intuitive conceptual basis for the most analogically productive dimensions, but there is an expectation that there will be*

JB-7: The condition (a-b)-(c-d)=0 is equivalent to (a-c)-(b-d)=0, making A:B::C:D equivalent to A:C::B:D from the point of view of a method of yours. Comment on the reasonableness or otherwise of this, and if possible on relevant evdience from psychological experiments.

> *I've added some additional analysis of the oblong geometry of analogies, which I think speaks to this valid point: there are two different ways to cut the conceptual transfer happening in an analogy, but these cuts are often quantifiably different. I've done this in the context of the potential for identifying equivalence between the mappings, and also included reference to relevant work from Tversky and Ortony.*

JB-8: p.153 para 1: As mentioned in the viva, you seem to be saying that, for each analogy separately, you find the best "top ... projection". This apperenlty conflicts with the 90% claim in next para, which seems to say there's a good space that works for 90% of the analogies. Apply the clarification that you gave in the viva.

> *I've clarified that I select the dimensions that most closely satisfy the equality between the word-vectors implied by the analogy, and then see if the analogy is in fact satisfactorily mapped in the resulting subspace.*

JB-9: Say whether you have looked at the vectors a-b, a-c etc to see how similar they might be to vectors for the expected underlying relationships in the analogy, and/or mention something about this in future work. Of course, if you have an argument for not expecting those difference vectors to be interesting, add this.

> *I've added a paragraph considering the interpretability of the vectors the describe the structure of an analogy, including a point about one potential problem (negative values), and also a suggestion that this points towards potentially productive future work.*

Miscellaneous

JB-10: For non-speclialist readers, state somewhere what precision and recall are and how f-score and accuracy are defined. You thesis may well be read by people who are not computational linguists or similar. The information could go in glossary at beginning, along with the other comparably basic concepts you include there.

> *I've added these definitions to the glossary.*

JB-11: Make sure it's clear, for each experiment, what the definitions of precision and recall are exactly in the particular case. Sometimes it's not clear what you're counting.

> *I've indicated the definition of coercion for both of the classification tasks (metaphor and coercion).*

JB-12: Glossary would also be good place to say that e.g. by a 2x2 co-occurrence window you mean one that goes two before and two after the target word.

> *I've updated the definition of co-occurrence window accordingly.*

JB-13: And after p.50: you actually talk about window size in terms of "k", so on first use of "NxN" for some N, point out you mean k=N.

> *I've clarified this both where I describe the parameter k and at the first mention of an NxN window.*

JB-14: p.41 and Fig 3.2: clarify that "co-occurrence terms" are selected words to which other words are being related. It only becomes gradually clear up to p.46 that you're talking about co-occurrence with "soprano" etc.

> *I've clarified that co-occurrence terms correspond to the dimensions selected for the analysis of a set of word-vectors, and have likewise clarified that the words in Figure 3.2 are the labels of co-occurrence dimensions selected for such an analysis.*

JB-15: p.49: I don't understand the point about parentheticals - add extra explanation.

> *I've clarified that parenthetical phrases often serve to interupt the sytagmatic flow of a sentence and therefore can skew the information available through co-occurrence counts.*

****

JB-16: p.51 para 1: The meaning given for the ratio (without the "a" correction) is wrong. The ratio can be bigger than 1 (right?) so cannot be any sort of probability! I think you mean: the denominator is proportional to the joint probability. The formula given on p.69 and p.149 in terms of probabilites makes the matter much clearer, and shows that the explanation on p.51 is askew. *****************

Yes, absolutely, this was a mistake: the ratio (omitting the constant "a") is the joint probabaility divided by the product of the independent probabilities. This has been corrected, and the more detailed analyses later and this chapter and then in Chapter 7 have been cross-referenced.

JB-17: Around here say something about how the chosen value of 10,000 for "a" comes about. What was this choice influenced by? Would it change of the corpus-size changed?? Etc.

I've noted that the value was determined through trial and error, but also corresponds approximately with the mean frequency of vocabulary words (ie, the 200,000 most frequent words. I've also noted that the value should scale linearly with corpus size.

***

JB-18: Explain at some early point that the "dimensions" are just the words along the top of the matrix M.

JB-19: p.53 (and as necessary elsewhere): You say "component" of T, but "member" and "element" would be clearer and standard. It's just a set, right?

I've amended this usaged, and a comprehensive search of the thesis doesn't turn up any other instances.

JB-20: In INDY: what about different components of T selecting some of the SAME dimensions, so you can't just concatenate the lists of dimensions found? You seem to assume there'll be no such duplication. Clarify this matter.

Yes, this was an omission: as with the ZIPPED technique, dimensions are selected iteratively and such that no dimension is selected more than once; I've made this clear.

JB-21: What does "top" amount to here?

By top I meant highest PMI value. I've made this clear.

JB-22: What if d/|T| isn't an integer?

The model iterates through the word-vectors in T until d is satisfied. In principal, this means there could be a small imbalance in dimensions associated with a given input, but in practice most of my experiments

> *explore dimensionalities that are divisible by the number of input vectors. I've clarified this in the thesis.*

JB-23: In explanation of JOINT: what does "merged" mean here? Multiplied or minimized point-wise? Becomes clear later, but it should be made clear here (in ordinary English, probably, rather than mathematically).

> *I've changed this to read "extracted from the base matrix M and aligned dimension-for-dimension".*

JB-24: Surely you discard the dimensions with *zero* elements, not non-zero elements!

> *Yes! Corrected.*

JB-25: For definiteness, explain that by normalised you mean in the standard vector-space sense of converting to length 1. "Normalise" means lots of different things in different areas.

> *I now described the truncated vectors as being "subjected to L2 normalisation, such that each vector is of uniform length".*

JB-26: p.54, para before formula 4.2: being "central" sounds contradictory with being "peripheral" - explain.

> *What I meant was "central and far from the origin". I've made this clear.*

JB-27: p.54: Formula 4.2 needs amendment. For one thing you seem to be using the curly brackets normally used for *set* notation to define a vector instead.

And anyway the "i" has no reference, but needs somehow to be tied to "t". Minimally "t" should be replaced by "t_h,i" (where my underscore means subscripting).

I think that what you want to is actually quite tricky to write in a clear formula, and it may be best to stick with saying that n_h is formed successively from those successive components t_h,i of t_h where ((... your Pi ¿ 0 condition as written in 4.2))

> *Yes, I see what you're saying, this is tricky to express. I've removed equation 4.2 altogether in favour of a sentential description, using your suggested nomenclature. (I've also changed the name of the normalised matrix from M to N', so as not to confuse it with the established convention for the base matrix.)*

JB-28: Formula 4.4: I don't understand the "d/k" subscript of "max".

> *This is also tricky. I'm trying to say that for each input, the top d/k dimensions are chosen from the corresponding word-vector. Perhaps this is also something that should only be expressed sententially, but I've added a little text and tried to rewrite the formula.*

JB-29: Somewhere up to here: I felt the need for some idea of typical values of d and —T—.

> *I've added a preview of expected values right after the introduction of the three subspace selection techniques, alongside my response to comment AK-3 below.*

***

JB-30: p.56, Table 5A and text: give more systematic information about statistical significance of the results.

JB-31: p.77: WordNet experiment re "body part" etc.: you mention "accuracy" but your explanations sound as though you're only looking at number of correct items found, i.e. measuring *recall*. (This contrasts with use of f-scores at other places, e.g. on p.114 bottom.) Make the text clearer.

> *This is a good and also subtle point: this is, in fact, a classification task, where the words in the intersection between the WordNet and model vocabulary are being labelled as in or out of a given category. As you say, I am not measuring accuracy; this would be besides the point, as there is a very large majority class of terms not included for each category. I've made this clear in the text, but just with the caveat that, by this analysis, I think these results are properly described as precision scores, since it is the number of correct classifications made by the model compared to the total number of classification made by the model (I don't think was clear in the text, but hopefully now is). Recall, and therefore f-score, would be misrepresentative in the cases where the models only return 50 terms, as these statistics would necessarily be very low for categories with much larger memberships. I've included a footnote to this effect.*

JB-32: Somewhere up to p.93: Optional amendment: I suggest that you make more of the superiority of your method over static methods on the lower dimensionalities.

> *I've added a short paragraph noting and briefly analysing this.*

JB-33: p.118: what agreement or whatever is the kappa measure measuring in this analysis?

> *It measures the difference of the accuracy and the probability of chance agreement divided by the compliment of the probability of chance agreement, so it's effectively a more stringent statistic as accuracy increases. I've added a couple lines about this.*

JB-34: p.122: end of middle para, relating to Fig 6.2: ”remain fairly far”: doesn't seem to be borne out by the Figure.

> *I've tempered the language; there's a bit of a trick of 3-D representation at play here, but it's very hard to get just the right perspective on these figures and I think the way they're presented is about as good as possible.*

JB-35: p.155: ”the analysis” – but that used for the prior results involved knowing the fourth term, no?

> *The subspaces projected in the prior proof-of-concept results were 20 to 200 dimensional subspaces projected based on knowledge of the fourth component of an analogy, selected from a set of 400 dimensions based only on the first three components. The purpose of this methodology was to establish that there was some way to cut a relatively small set of dimensions based on a naive analysis of the analogy. I've made this more clear at this point in the text.*

Important Typos etc.

NB there are quite a few minor typos spread over the thesis - it needs a further round of complete proofreading.

> *Thank you for pointing these out. I've corrected all of them, and many others. I haven't used my correction formatting for typos, as I'm sure none of the updates will be controversial and I don't want to distract from the more substantial revisions.*

Some of the more important things to fix:

p.25 near top: do you really mean ”asseveration”? Just mean ”assertion”? Your word sounds a bit judgmental.

p.32 first full para: criteria -¿ criterion p.61, Table 4C: "angel" !

p.66, l.-5: distances -¿ distances between {I presume}

p.126, p.137 l.5, p.170 l.-6, p.171 l.2: discreet -¿ discrete {discreet has a completely different meaning!}

p.138 top: content -¿ context

2 Anna Korhonen's Comments and Responses

General corrections:

AK-1: Explain, for all the experiments, whether you calculated statistical significance and whether the improvements or differences reported are significant.

> *I've clarified that probabilities of chance observation of p ¡ .01, based on the tests outlined in Section 1.3, will be considered statistically significant, and that comparisons are made between model techniques with similar dimensionality parameters. I've stated this in Section 1.3 as well as in the background sections for each experiment. I realise there is more to say about hypothesis testing in NLP, and have included references to further reading in Section 1.3.*

AK-2: Explain, for all the experiments to which this applies, how you did qualitative analysis: did you analyse all of the output data or only a subset? Did you do the analysis in some systematic manner?

> *I've added a titled paragraph to Section 1.3 explaining that qualitative analysis will be performed by selecting examples from the extents of model output, as well as based on other patterns in the data in, for instance, the case of the two-axes plots of relatedness/similarity scores. I've also added specific details about qualitative analysis as it arises in each experiment.*

AK-3: Explain how you came up with the thresholds, the number of dimensions (20, 50, 200 etc) and sizes of windows used in your experiments. Did you just invent them or are they based on some preliminary experiment and then adopted for the rest of the thesis?

> *The thresholds for dimensionality are intended to be a sample up to the point where the joint technique becomes problematic due to a lack of mutually non-zero dimensions for some inputs. The co-occurrence window sizes are representative of two very common choices as reported in the literature. I've indicated this in the text (in the first case in conjunction with comment JB-29 above).*

AK-4: This is an optional correction, but I would recommend highlighting the best results in the tables for easier readability.

> *Done! (For instances where I compare categorically across different parameters; for instances where I compare singular results from other studies, scores are generally listed in order.)*

Specific corrections:

AK-5: p. iv Glossary: Because the terminology is quite cross-disciplinary in this thesis, I would expand this glossary a bit, give a more comprehensive definition for each concept and even mention, for concepts that are used in a non-conventional (from NLP perspective) sense, which field the definition comes from. I would also add meaning and situation(al) in this glossary.

Chapter 1

AK-6: Explain more clearly that this thesis belongs to / advances primarily the field of computational linguistics.

AK-7: Define more clearly what the thesis does and doesnt do: it develops computational linguistic methodology and evaluates it in the context of NLP tasks. Although based on theoretical insights from other fields (and although potentially useful for several fields), cross-disciplinary investigations are not included but are left for future work.

AK-8: P. 10 typo: there are two the words in the 3rd line of the 3rd paragraph

Chapter 2

AK-9: Section 2.2. When you start talking about concepts here, please define the intended properly and/or refer to the places in the thesis where they are defined properly (and mention the existence of that Glossary).

AK-10: Section 2.3 I found all this background on metaphor was a little out of place in this section. Consider moving the background elsewhere if possible?

AK-11: p. 24 This is the place where I would mention that no attempt is made in this thesis to discuss the conceptual structure (of the human brain).

AK-12: Section 2.4. Show some awareness of the long history of semantics research in NLP by including a paragraph or two that mention the main lines of research prior to / alongside VSMs. Then say explicitly that you will focus your literature review literature on the distributional semantics and neural approaches only.

AK-13: On p. 28 define generalizability better (its a term with many meanings).

Chapter 3

AK-14: Section 3.2. p. 37 The concept situation should be defined better. In particular, what does a situation of words in a large corpus mean?

AK-15: p. 38 Context is defined here, too late in the write-up.

Chapter 4

AK-16: Section 4.1 The second paragraph talks about the cleaning process. Mention who performed this process.

AK-17: Section 4.2 Theres a typo in the caption of table 4-7: delete one the

Chapter 5

AK-18: Here or earlier: Explain how you chose the specific tasks in chapters 5, 6, and 7 for your evaluation?

AK-19: p. 98 Here or elsewhere in the chapter (or in the future work section at the end of the thesis) discuss how factors such as part-of-speech (nouns, verbs, adjectives), polysemy and abstract vs. concrete, among others, many also influence your results alongside the obvious issue of frequency.

AK-20: P. 130 The end of the first paragraph on this page is difficult to understand explain what you mean by too conventional

Chapter 6

AK-21: p. 139 BNC is a balanced corpus so shouldnt be colloquial in nature. Explain, if you can, how the 2000 sentences were selected.

Chapter 8

AK-22: Section 8.2. If possible, I would try and discuss the potential usefulness of the methodology introduced in this thesis for NLP at large, and for real-life applications (search, QA, etc). I would also discuss what it would take to make your methods useful for research in cognitive science.