Lang Resources & Evaluation (2008) 42:99–107
DOI 10.1007/s10579-008-9063-y

Dimensionality of dialogue act tagsets

An empirical analysis of large corpora

Andrei Popescu-Belis

Published online: 22 February 2008

Abstract This article compares one-dimensional and multi-dimensional dialogue act tagsets used for automatic labeling of utterances. The influence of tagset dimensionality on tagging accuracy is first discussed theoretically, then based on empirical data from human and automatic annotations of large scale resources, using four existing tagsets: DAMSL, SWBD-DAMSL, ICSI-MRDA and MALTUS. The Dominant Function Approximation proposes that automatic dialogue act taggers could focus initially on finding the main dialogue function of each utterance, which is empirically acceptable and has significant practical relevance.

Keywords Dialogue act tagsets · Conversational corpora · Tagset dimensionality

1 Introduction

The communicative functions of utterances in dialogue-based interactions are often called *dialogue acts* (DAs). This article attempts to frame a central question formulated by Traum (2000): should one-dimensional DA tagsets be preferred over multi-dimensional ones, or vice-versa? The main factor that will be considered here is the size of the search space for a human annotator or an automatic tagger, i.e. the number of possible labels to choose from (Sect. 2).

As smaller search spaces facilitate classification, one-dimensional tagsets have an empirical advantage over multi-dimensional ones, exemplified here on four tagsets (Sects. 3–5). However, the theoretical bases of dialogue acts suggest that utterances have multi-dimensional functions (Popescu-Belis 2005, Sect. 3), a property that is

A. Popescu-Belis (✉)
IDIAP Research Institute, Centre du Parc, Av. des Prés-Beudin 20,
P.O. Box 592, 1920 Martigny, Switzerland
e-mail: andrei.popescu-belis@idiap.ch

better captured by a multi-dimensional scheme such as those introduced by Traum and Hinkelman (1992) or by Bunt (2005).

The compromise proposed here, called the Dominant Function Approximation, consists in defining a multi-dimensional tagset with one default function in each of its dimensions, then focusing manual and automatic annotation on the single most important non-default function of an utterance. In agreement with current practice, the dimensions should capture aspects of the functions of utterances—not always directly related to their form—based on their illocutionary effects (direct or indirect speech acts) or on strong implicatures. However, as perlocutionary effects or weak implicatures are more difficult to grasp using a finite tagset, they require a more elaborate dialogue annotation model.

2 Tagset dimensionality

In the mathematical sense, *dimensions* are the parameters required to describe the DA labels of utterances, and are characterized by sets of possible values. A one-dimensional tagset is a set $A = \{a_1, a_2, \ldots\}$, where each utterance is tagged with exactly one elementary tag $a_n \in A$. A multi-dimensional tagset is a set of dimensions $\mathcal{T} = \{A, B, \ldots\}$ where each dimension is in turn a list of tags, $A = \{a_1, a_2, \ldots\}$, $B = \{b_1, b_2, \ldots\}$ and so on. Each utterance is then tagged with a composite label or tuple of tags $(a_i, b_j, \ldots)$, i.e. exactly one tag from each dimension.

In case the annotation guidelines allow the use, for each utterance, of all tags that may apply from a set A, the set should not be called a dimension in the sense described above, as the tags are not mutually exclusive. Instead, the resulting tagset is equivalent to the following Cartesian product of proper dimensions: $\{a_1, \emptyset\} \times \{a_2, \emptyset\} \times \ldots$, which describe all possible combinations of tags from A.

The *dimensionality* of a tagset is the number of its proper dimensions, that is, sets containing mutually-exclusive tags. If a multi-dimensional tagset has N dimensions, each of size k_i, then the size of the tagset (the number of possible labels) is $k_1 \times k_2 \times \cdots \times k_N$, a potentially very large number. Moreover, if these are not proper dimensions, i.e. if any number of tags from each set can be applied to an utterance, then the number of possible labels becomes $2^{k_1} \times 2^{k_2} \times \cdots \times 2^{k_N}$, an even larger number.

3 Theoretical vs observed DA labels for four tagsets

Four DA tagsets are selected here to illustrate the difficulty of choosing between a one-dimensional and a multi-dimensional tagset. Compared to the numerous other DA tagsets that have been proposed, these are among the few general-domain tagsets that have been used to annotate large scale resources.

The Dialogue Act Markup in Several Layers, DAMSL (Core and Allen 1997), distinguishes four aspects of utterance-function, which are not proper dimensions, as an utterance may be tagged with as many tags as needed from each set. The application of DAMSL to the Switchboard corpus led to the development of the one-dimensional SWBD-DAMSL tagset with 42 tags, attempting to reduce DAMSL's dimensionality, in particular for automatic DA tagging (Jurafsky et al. 1998). While the number of possible combinations of DAMSL tags is about 4 million (Clark and Popescu-Belis 2004), Jurafsky et al. (1998) observed that only 220 different ones occurred in the 200,000 utterances of the Switchboard corpus. These were further merged into 42 mutually-exclusive, synthetic SWBD-DAMSL tags, which stand for specific combinations of elementary DAMSL functions. Therefore, although the number of tags in SWBD-DAMSL is nearly the same as in DAMSL, the fact that no combination of SWBD-DAMSL tags is allowed results in a considerably smaller search space for an automatic tagger.

The annotation of the ICSI Meeting Recorder corpus (ICSI-MR) allowed again the combination of as many SWBD-DAMSL-style tags as needed for each utterance (Shriberg et al. 2004). The resulting multi-dimensional tagset, ICSI-MRDA, thus removed SWBD-DAMSL's mutual-exclusiveness constraint. Although in theory ICSI-MRDA placed no restriction on the number of tags per utterance, in practice annotators have used up to six tags per utterance for the ICSI-MR corpus.

The number of possible ICSI-MRDA labels that have at most six tags reaches several million, as shown in Table 1 below. The number of possible *types* of labels increases exponentially with the number of tags per label, while the number of observed types first increases, then decreases to zero. It appears from the fourth column of Table 1 that only 776 different types of ICSI-MRDA labels occur for the 113,560 utterances of the ICSI-MR corpus. Among these, 69% of the types and 98% of the tokens are composed of 1, 2 or 3 tags.

The rightmost column of Table 1 shows the maximal tagging accuracy that can potentially be reached on the ICSI-MR corpus using an ideal tagger (an oracle) that is limited to ICSI-MRDA labels with at most *N* tags. For instance, if only labels made of 1, 2 or 3 tags were used for automatic tagging, then only about 2% of the utterances would be intrinsically impossible to tag correctly, but the search space would be reduced from several million combinations to only 8,591 labels.

Table 1 Number of theoretical and observed ICSI-MRDA and MALTUS labels (types and tokens) on the ICSI-MR corpus with 113,560 utterances. The last column gives the accuracy of an oracle tagger limited to at most *N* tags per label

Tagset	Tags/ label	Possible label types	Observed label types	Observed label tokens	Max. acc.
ICSI-MRDA	1	11	11	68,213	0.6007
	2	429	129	37,889	0.9343
	3	8,151	402	5,054	0.9788
	4	100,529	176	2,064	0.9970
	5	904,761	49	326	0.9999
	6	6,333,327	9	14	1.0000
MALTUS	1	4	4	84,092	0.74051
	2	28	14	28,366	0.99003
	3	72	29	1,089	0.99997
	4	88	3	3	1.00000

The MALTUS tagset (Multidimensional Abstract Layered Tagset for Utterances) was introduced in order to reduce the number of possible dimensions and labels. MALTUS merged some of the ICSI-MRDA tags, and grouped tags into classes by hypothesizing mutual-exclusiveness constraints (Clark and Popescu-Belis 2004). MALTUS has six dimensions, four being binary ones, and therefore its size is several orders of magnitude smaller than that of DAMSL or ICSI-MRDA tagsets. There are indeed no more than 600 possible MALTUS labels, or only 192 if disruptions (i.e. unfinished or interrupted utterances) are not considered, as shown in the lower part of Table 1. Only 50 MALTUS labels appear in the 113,560 utterances of the ICSI-MR corpus, once ICSI-MRDA is converted to MALTUS. A specific count shows that only 22 MALTUS labels occur more than 20 times each, with the 6 that appear more than 5,000 times each being S (statement) 51,304 times, B (backchannel) 15,180 times, H (floor-holder) 12,288 times, S^AT (attention-related statement) 8,280 times, S^RP (positive response) 7,612 times, and Q (question) 5,320 times.

As was the case for ICSI-MRDA, if only the 22 most frequent MALTUS labels were used for automatic tagging of the ICSI-MR corpus, then only 0.12% of the utterances (136 out of 113,560) would be impossible to tag correctly, while the search space would be reduced from 192 to 22 labels. In other words, an oracle tagger limited to the 22 most frequent MALTUS labels could reach 99.88% accuracy on ICSI-MR, and, presumably, a similar value on comparable data. This is well above the actual performances of automatic taggers or the observed agreement of human annotators, and suggests considering for a start only the reduced subset of tags.

4 Effects of dimensionality on manual tagging

Few experiments analyze directly the impact on inter-annotator agreement of tagset size, as tagsets are often fixed from the beginning of a project. Comparisons using different DA tagsets on the same data are costly and therefore infrequent (Carletta et al. 1997). While acknowledging the limits of comparisons over different data, this section gathers some of the agreement scores available in the literature, showing that agreement tends to decrease when the size of the tagset increases.

Inter-annotator agreement, often measured by the *kappa* score (Di Eugenio and Glass 2004), is generally considered to be good when $\kappa > 0.8$, and acceptable when $\kappa > 0.67$. Such values often characterize low-dimension tagsets: for instance, $\kappa = 0.8$ for SWBD-DAMSL (Jurafsky et al. 1998). To reach the same value on the ICSI-MR corpus, Shriberg et al. (2004, p. 99) applied a class map reducing the ICSI-MRDA tagset to only five abstract labels: statement, question, turn management, backchannel, and disruption. Using a more detailed class map with about 15 labels, κ decreased to 0.76. In an experiment with a one-dimensional tagset, Doran et al. (2003, p. 136) found that inter-annotator agreement decreased when the size of the tagset increased: $\kappa = 0.90$ for 20 tags, but $\kappa = 0.71$ for 26 tags. This surprisingly large decrease could also be explained by an adequacy problem: the larger tagset could have been less adapted to the targeted phenomena, and therefore more difficult to apply.

Di Eugenio et al. (2000) studied inter-annotator agreement on about 500 utterances tagged with a DAMSL-inspired tagset, and found that κ varied from 0.83 to

0.54 for various dimensions, often in proportion to their size, among other factors. For instance, $\kappa = 0.79$ for the {answer, $\varnothing$} binary dimension; $\kappa = 0.72$ for the {offer, commit, $\varnothing$} dimension; and $\kappa = 0.54$ for {accept, reject, hold, $\varnothing$}. These figures are slightly higher than those obtained by Core and Allen (1997, Tables 2 and 3), probably due to a better adaptation of the annotation guidelines to the type of data.

In an experiment with the DIT++ tagset, Geertzen and Bunt (2006) found that the highest value of κ, 0.82, was reached for the turn management dimension with only four tags. For the task-related dimension, with more than 40 general-purpose tags, κ is 0.47, a much lower value, which can be improved to 0.71 only if the comparison metric takes into account similarity of tags, which is an indirect way to reduce the tagset size. The agreement for the auto-feedback dimension is even lower at $\kappa = 0.21$ (corrected at 0.57), and this dimension has even more combinations of tags than the previous one.

The recent dialogue act annotation of the AMI Meeting Corpus used a new one-dimensional tagset with 16 tags. The results of pilot studies of inter-annotator agreement were acceptable enough to allow large scale annotation (139 scenario-based meetings with 117,887 utterances), but no value has yet been published. Here, again, acceptable inter-annotator agreement is related to a one-dimensional tagset with less than 20 labels.

5 Effects of dimensionality on automatic tagging

The influence of tagset dimensionality on automatic DA tagging is quite complex to assess *a priori*. An automatic tagger for a multi-dimensional tagset can consist of separate classifiers for each dimension, or of a joint classifier over combinations of tags, while for a one-dimensional tagset it is natural to use a unique classifier (each class corresponding to one of the mutually-exclusive tags). For multi-dimensional tagsets, a joint classifier might seem preferable as its search space tends to be much smaller than that of separate classifiers, because the classifier can learn from the data the joint probability distribution of the combinations of tags, including for instance information about unlikely combinations. However, dimension-specific classifiers could also be preferable, because the data they process (a subset of the full annotation) has fewer degrees of freedom, and therefore such classifiers require fewer features for classification than joint ones, which increases their accuracy if the amount of training data is constant.

In the comparative overview provided by Samuel (1999, p. 29) the accuracy of automatic DA tagging for various methods and tagsets is between 46% and 75%. However, the role of tagset dimensionality is blurred by the even larger influence of the training data and the nature of the classifiers. If all other things were equal, a smaller number of classes facilitates statistical classification, a fact illustrated for instance in an experiment with 220 vs. 42 DA tags (Webb et al. 2005).

Clark and Popescu-Belis (2004) showed that DA tagging using separate classifiers for each dimension of the MALTUS tagset had lower accuracy than a single, combined classifier. The joint classifier scored 73.2% on the ICSI-MRDA data, compared to 70.5% reached by combined classifiers, all other things being equal. It is likely that the joint classifier performed better than the independent classifiers because the latter could not model obvious dependencies between dimensions—in other words, it could not see that only a very small fraction of all possible combinations of tags really occurred in the data. If confirmed by other experiments, these results show that, below asymptotic performance, a reduced search space is a more effective way to increase accuracy than looking at each dimension separately, which offers proportionally more data but does not capture the dependencies between dimensions.

The figures and analyses from Sects. 3–5 suggest that smaller tagsets tend to lead to higher human and automatic performance for DA annotations. In addition, one of the main reasons why some tagsets are particularly large is the multiplicative factor described in Sect. 2: $k_1 \times k_2 \times \cdots \times k_N$, the size of a multi-dimensional tagset, is much larger than $k_1 + k_2 + \cdots + k_N$, the size of a one-dimensional tagset with the same number of tags. Therefore, solutions to the "curse of dimensionality" should try to reduce first the size of the search space. It is thus preferable to define the dimensions of a tagset as sets of mutually-exclusive tags, and also to avoid multiplying them beyond necessity. Constraints across dimensions should be found whenever possible, based on theoretical and empirical evidence, so that the combinations of tags that cannot occur can be explicitly ruled out. This can be done either by removing them from a linearized version of the tagset (obtained by enumerating all possible combinations of tags) or by developing more advanced classification models that handle cross-dimensional constraints.

6 The Dominant Function Approximation

The Dominant Function Approximation (DFA) is a more principled way to deal with tagset dimensionality. Starting with a theoretically-motivated multi-dimensional tagset, along the dimensions outlined in (Popescu-Belis 2005, Sect. 3), fixed default functions should be identified in each dimension, based on linguistic and pragmatic considerations (e.g. "unmarked" utterance functions) or set a posteriori based on frequency counts. The DFA hypothesizes that *every utterance has only one main communicative function*, called its *dominant* function, and that its functions in all the other dimensions are the default ones. (It is also possible that all functions of an utterance are the default ones.) In other words, speakers most often accomplish only one non-default dialogue act at a time.

For instance, turn-taking is often managed by implicit cues, and most utterances will have the unmarked (default) role in this functional dimension: take the turn initially and release it at the end of the utterance. However, when turn-taking must be managed explicitly, the role of an utterance becomes dominant in the turn-taking dimension and unmarked (default) in all the other ones. For instance, uttering 'But…' to take one's turn does not express opposition or feedback. Similarly, politeness is constantly managed in a dialogue, but when utterances fulfill an explicit function in this dimension (e.g. greetings), they are often also limited to it, as acknowledged for instance in the AMI Project guidelines (AMI Project 2005, Sect. 3.5): “Only classify an act as social if it does not also fit in one of the other groups.”

The DFA is a working hypothesis which states that the functional description of utterances can be slightly simplified in order to facilitate automatic and manual annotation. The DFA effectively transforms a multi-dimensional tagset into a one-dimensional one, whose size is the sum of the sizes of the original dimensions rather than their product. Automatic DA taggers looking only for the dominant function of each utterance would therefore benefit from this dramatic reduction of the search space, which potentially increases their accuracy. The DFA is applicable to language technology, as focusing only on the literal functions of utterances might be acceptable for current human-computer dialogue systems, though it could be less acceptable for a precise analysis of human dialogues. When generating an utterance, conveying two functions in two successive utterances is an acceptable solution, though it might be less efficient than combining them into a single one.

To study the DFA empirically, DA annotators could be asked to identify only the dominant function of each utterance, for instance by indicating first the dominant dimension (in this case, the perceived size of the tagset is the number of its dimensions plus the size of the largest dimension): high agreement scores would support the DFA. Alternatively, annotators could assign to each utterance as many functions as necessary, and the DFA would be supported if the proportion of utterances with more than one non-default function was reasonably small compared to the typical uncertainty of the annotation process.

The latter test can be applied to existing annotations as well. For instance, using the MALTUS tagset on the ICSI-MR corpus, if S (statement) is considered to be the default function in the general-level dimension, and ‘null’ the default function in all the other dimensions, then all utterances labeled only with a general tag, or with S plus only one specific tag, satisfy the DFA. There are indeed 97% such utterances in the ICSI-MR corpus (Popescu-Belis 2005, p. 20), the main exception being tag questions (2.7%). To derive similar figures for ICSI-MRDA, it is necessary to organize first its 54 tags into proper dimensions and to define default tags, but a quick count of all labels composed of only one first-tier tag, or of ‘statement’ and one second-tier tag shows that at least 92% of the labels satisfy the DFA. In the case of one-dimensional tagsets, most of the SWBD-DAMSL composite tags reflect one dominant function in one dimension, as do the AMI Project tags. In these cases, the high inter-annotator agreement values offer support for the DFA.

7 Conclusion

This paper has identified two opposing factors that influence the definition of DA tagsets. Multi-dimensional tagsets appear to find theoretical justifications in the multiplicity of functions that utterances can fulfill, but they also have search spaces that are several orders of magnitude higher than those of one-dimensional tagsets, a fact that tends to decrease the accuracy of human and automatic annotations. Multi-dimensional DA tagsets that are inspired by dialogue theories and are accompanied by the Dominant Function Approximation could get the best of both worlds: their theoretical basis facilitates understandability and interoperability, while their use in computational applications as a one-dimensional tagset, thanks to the DFA, is likely to increase the accuracy of DA recognition. Further empirical results could help assess the margin of error of the Dominant Function Approximation and extend this hypothesis to other multi-dimensional tagging problems, such as subjectivity or argumentation.

Acknowledgements This work has been supported by the Swiss National Science Foundation within the IM2 NCCR on Interactive Multimodal Information Management (www.im2.ch). The author would like to thank Jean Carletta, Hannah Carvey, Alex Clark, Liz Shriberg, Sandrine Zufferey, and the anonymous reviewers for their helpful comments on this paper.

References

AMI Project. (2005). Guidelines for dialogue act and addressee annotation. Augmented Multiparty Interaction Project Document, v. 1.0, 13 October 2005. http://www.corpus.amiproject.org.

Bunt, H. (2000). Dynamic interpretation and dialogue theory. In M. M. Taylor, F. Néel, & D. G. Bouwhuis (Eds.), *The structure of multimodal dialogue II* (pp. 139–166). Amsterdam: John Benjamins.

Bunt, H. (2005). A framework for dialogue act specification. In *Fourth Workshop on Multimodal Semantic Representation*. Tilburg.

Carletta, J., Isard, A., Isard, S., Kowtko, J. C., Doherty-Sneddon, G., & Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics, 23*(1), 13–32.

Clark, A., & Popescu-Belis, A. (2004). Multi-level dialogue act tags. In *Fifth SIGdial Workshop on Discourse and Dialogue* (pp. 163–170). Cambridge, MA.

Core, M. G., & Allen, J. F. (1997). Coding dialogues with the DAMSL annotation scheme. In D. R. Traum (Ed.), *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines* (pp. 28–35). Menlo Park, CA.

Di Eugenio, B., & Glass, M. (2004). The kappa statistic: A second look. *Computational Linguistics, 30*(1), 95–101.

Di Eugenio, B., Jordan, P. W., Thomason, R. H., & Moore, J. D. (2000). The agreement process: An empirical investigation of human-human computer-mediated collaborative dialogues. *International Journal of Human Computer Studies, 53*(6), 1017–1076.

Doran, C., Aberdeen, J., Damianos, L., & Hirschman, L. (2003). Comparing several aspects of human-computer and human-human dialogues. In J. van Kuppevelt & R. W. Smith (Eds.), *Current and new directions in discourse and dialogue* (pp. 133–159). Dordrecht: Kluwer.

Geertzen, J., & Bunt, H. (2006). Measuring annotator agreement in a complex hierarchical dialogue act annotation scheme. In *Seventh SIGdial Workshop on Discourse and Dialogue* (pp. 126–133). Sydney.

Jurafsky, D., Shriberg, E., Fox, B., & Curl, T. (1998). Lexical, prosodic, and syntactic cues for dialog acts. In *Coling-ACL 1998 Workshop on Discourse Relations and Discourse Markers* (pp. 114–120). Montreal.

Lesch, S., Kleinbauer, T., & Alexandersson, J. (2005). Towards a decent recognition rate for the automatic classification of a multidimensional dialogue act tagset. In *Fourth Workshop on Knowledge and Reasoning in Practical Dialogue Systems* (pp. 46–53). Edinburgh.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press.

Popescu-Belis, A. (2005). Dialogue act tagsets: One or more dimensions? *ISSCO Working Paper 62*, University of Geneva.

Samuel, K. (1999). Discourse learning: An investigation of dialogue act tagging using transformation-based learning. Ph.D. thesis, University of Delaware, Department of Computer and Information Sciences.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., & Carvey, H. (2004). The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Fifth SIGdial Workshop on Discourse and Dialogue* (pp. 97–100). Cambridge, MA.

Traum, D. R. (2000). 20 Questions for dialogue act taxonomies. *Journal of Semantics, 17*(1), 7–30.

Traum, D. R., & Hinkelman, E. A. (1992). Conversation acts in task-oriented spoken dialogue. *Computational Intelligence, 8*(3), 575–599.

Webb, N., Hepple, M., & Wilks, Y. (2005). Dialogue act classification based on intra-utterance features. In *Fifth AAAI Workshop on Spoken Language Understanding*. Pittsburgh, PA.