Querent Intent in Multi-Sentence Questions

Laurie Burchell[*], **Jie Chi**[*], **Tom Hosking**[*], **Nina Markl**[*], **Bonnie Webber**
Institute for Language, Cognition and Computation
University of Edinburgh
{laurie.burchell,jie.chi,tom.hosking,nina.markl}@ed.ac.uk

Abstract

Multi-sentence questions (MSQs) are sequences of questions connected by relations which, unlike sequences of standalone questions, need to be answered as a unit. Following Rhetorical Structure Theory (RST), we recognise that different "question discourse relations" between the subparts of MSQs reflect different speaker intents, and consequently elicit different answering strategies. Correctly identifying these relations is therefore a crucial step in automatically answering MSQs. We identify five different types of MSQs in English, and define five novel relations to describe them. We extract over 162,000 MSQs from Stack Exchange to enable future research. Finally, we implement a high-precision baseline classifier based on surface features.

1 Introduction

A multi-sentence question (MSQ) is a dialogue turn that contains more than one question (cf. Ex. (1)). We refer to the speaker of such a turn as a *querent* (i.e., one who seeks).

(1) Querent: How can I transport my cats if I am moving a long distance? (Q1)
For example, flying them from NYC to London? (Q2)

A standard question answering system might consider these questions separately:

(2) A1: You can take them in the car with you.
A2: British Airways fly from NYC to London.

However, this naïve approach does not result in a good answer, since the querent intends that an answer take both questions into account: in (1), Q2 clarifies that taking pets by car is not a relevant option. The querent is likely looking for an answer like (3):

(3) A: British Airways will let you fly pets from NYC to London.

Whilst question answering (QA) has received significant research attention in recent years (Joshi et al., 2017; Agrawal et al., 2017), there is little research to date on answering MSQs, despite their prevalence in English. Furthermore, existing QA datasets are not appropriate for the study of MSQs as they tend to be sequences of standalone questions constructed in relation to a text by crowdworkers (e.g. SQuAD (Rajpurkar et al., 2016)). We are not aware of any work that has attempted to improve QA performance on MSQs, despite the potential for obvious errors as in the example above.

Our contribution towards the broader research goal of automatically answering MSQs is as follows:

- We create a new dataset of 162,745 English two-question MSQs from Stack Exchange.
- We define five types of MSQ according to how they are intended to be answered, inferring intent from relations between them.
- We design a baseline classifier based on surface features.

[*]Alphabetical order, equal contribution

The 14th Linguistic Annotation Workshop, pages 138–147
Barcelona, Spain (Online), December 12, 2020.

2 Prior work

Prior work on QA has focused on either single questions contained within dialogue (Choi et al., 2018; Reddy et al., 2019; Saeidi et al., 2018; Clark et al., 2018), or questions composed of two or more sentences crowd-sourced by community QA (cQA) services (John and Kurian, 2011; Tamura et al., 2005). Our definition of MSQs is similar to the latter, but it should be noted that sentences in existing cQA datasets can be declarative or standalone, while in our case they must be a sequence of questions that jointly imply some user intent. Popular tasks on cQA have only considered the semantics of individual questions and answers, while we are more focused on interactions between questions.

Huang et al. (2008) and Krishnan et al. (2005) classify questions to improve QA performance, but their work is limited to standalone questions. Ciurca (2019) was the first to identify MSQs as a distinct phenomenon, and curated a small dataset consisting of 300 MSQs extracted from Yahoo Answers. However, this dataset is too small to enable significant progress on automatic classification of MSQ intent.

3 Large-scale MSQ dataset

Stack Exchange is a network of question-answering sites, where each site covers a particular topic. Questions on Stack Exchange are formatted to have a short title and then a longer body describing the question, meaning that it is far more likely to contain MSQs than other question answering sites, which tend to focus attention on the title with only a short amount of description after the title. There is a voting system which allows us to proxy well-formedness, since badly-formed questions are likely to be rated poorly. It covers a variety of topics, meaning that we can obtain questions from a variety of domains.

To obtain the data, we used the Stack Exchange Data Explorer[1], an open source tool for running arbitrary queries against public data from the Stack Exchange network. We chose 93 sites within the network, and queried each site for entries with at least two question marks in the body of the question. We removed any questions with TEX and mark-up tags, then replaced any text matching a RegEx pattern for a website with '`[website]`'. From this cleaned text, we extracted pairs of MSQs by splitting the cleaned body of the question into sentences, then finding two adjacent sentences ending in '`?`'. We removed questions under 5 or over 300 characters in length. Finally, we removed any question identified as non-English using `langid.py` (Lui and Baldwin, 2012). Many of the questions labelled as 'non-English' were in fact badly formed English, making language identification a useful pre-processing step.

After cleaning and processing, we extracted 162,745 questions from 93 topics[2]. A full list of topics and the number of questions extracted from each is given in Appendix A. We restrict the dataset to pairs of questions, leaving longer sequences of MSQs for future work.

4 MSQ type as a proxy for speaker intent

MSQs are distinct from sequences of standalone questions in that their subparts need to be considered as a unit (see (1) in Section 1). This is because they form a **discourse**: a coherent sequence of utterances (Hobbs, 1979). In declarative sentences, the relationship between their different parts is specified by "discourse relations" (Stede, 2011; Kehler, 2006), which may be signalled with discourse markers (e.g. *if*, *because*) or discourse adverbials (e.g. *as a result*, see Rohde et al. (2015)). We propose adapting the notion of discourse relations to interrogatives.

A particularly useful approach to discourse relations in the context of MSQs is Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), which understands them to be an expression of the speaker's communicative intent. Listeners can infer this intent under the assumptions that speakers are "cooperative" and keep their contributions as brief and relevant as possible (Grice, 1975). Transposing this theory to interrogatives, we can conceptualise the querent's communicative intent as a specific kind of answer. Reflecting this intent, the relation suggests an answering strategy.

We introduce five types of "question discourse relations" with a prototypical example from our data set, highlighting the inferred intent and the proposed answering strategy in Table 1.

[1] https://data.stackexchange.com/

[2] Our dataset is available at `https://github.com/laurieburchell/multi-sentence-questions`

SEPARABLE	
Example	*What's the recommended kitten food?* *How often should I feed it?*
Intent	Two questions on the same topic (the querent's kitten).
Strategy	Resolve coreference and answer both questions separately.

REFORMULATED	
Example	*Is Himalayan pink salt okay to use in fish tanks?* *I read that aquarium salt is good but would pink salt work?*
Intent	Speaker wants to paraphrase Q1 (perhaps for clarity).
Strategy	Answer one of the two questions.

DISJUNCTIVE	
Example	*Is it normal for my puppy to eat so quickly?* *Or should I take him to the vet?*
Intent	Querent offers two potential answers in the form of polar questions.
Strategy	Select one of the answers offered (e.g. "Yes, it is normal") or reject both (e.g. "Neither – try feeding it less but more often").

CONDITIONAL	
Example	*Has something changed that is making cats harder to buy?* *If so, what changed?*
Intent	Q2 only matters if the answer to Q1 is "yes".
Strategy	First consider what the answer to Q1 is and then answer Q2.

ELABORATIVE	
Example	*How can I transport my cats if I am moving a long distance?* *For example, flying them from NYC to London?*
Intent	Querent wants a more specific answer.
Strategy	Combine context and answer the second question only.

Table 1: The five types of MSQ we describe, an example of each, the querent's intent, and the resulting answering strategy. Mann and Thompson (1988)'s ELABORATION, CONDITION and RESTATEMENT relations correspond roughly to three of the relations we recognise.

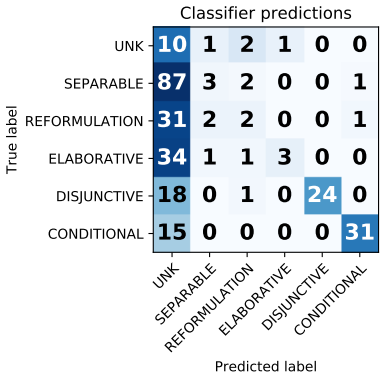
5 Classification using contrastive features

Since Ciurca (2019) found that using conventional discourse parsers created for declaratives is not suitable for extracting discourse relations from MSQs, we design our own annotation scheme and use it to implement a baseline classifier. Following previous work on extracting discourse relations (Rohde et al., 2015), we use discourse markers and discourse adverbials alongside other markers indicative of the structure of the question (listed in appendix C) to identify explicitly signalled relations.[3]

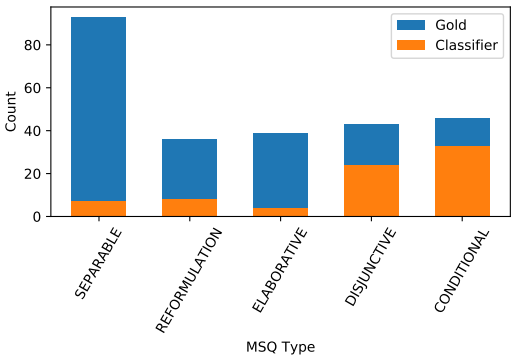
We construct a high-precision, low-recall set of rules to distinguish the most prototypical forms of the five types using combinations of binary contrastive features. To derive the relevant features, we consider the minimal edits to examples of MSQs required to break or change the type of discourse relation between their parts. We then define a feature mask for each MSQ type which denotes whether each feature is required, disallowed or ignored by that type. Each mask is mutually exclusive by design.

Given a pair of questions, the system enumerates the values of each feature, and compares to the definitions in Appendix B. If a match is found, the pair is assigned the corresponding MSQ label, otherwise

[3]Since implicit discourse relations are pervasive and challenging to automatic systems (Sporleder and Lascarides, 2008), we make no attempt to extract them here.



(a) Confusion matrix for our classifier evaluated on our hand-annotated test set. The classifier can reliably detect DISJUNCTIVE and CONDITIONAL MSQs, achieving a high overall precision score of 82.9%.



(b) Counts of each MSQ type in our test set, according to our annotation and our classifier. While SEPARABLE MSQs appear to be the most prevalent, the classifier identifies only a small fraction of them, implying that they are likely to be implicitly signalled. DISJUNCTIVE and CONDITIONAL are the most likely to be explicitly signalled.

Figure 1

it is assigned UNKNOWN. This process is illustrated in Appendix D.

To evaluate our classifier, 420 MSQs from our test set were annotated by two native speakers. We then evaluate the classifier on the subset of 271 samples for which both annotators agreed on the MSQ type. The resultant confusion matrix is shown in Figure 1a, with the classifier achieving 82.9% precision and 26.5% recall.

Overall, we find that our classifier performs well for a heuristic approach, but that real world data contains many subtleties that can break our assumptions. During the annotation process, we found many instances of single questions followed by a question which fulfils a purely social function, such as "*Is it just me or this a problem?*" (a *phatic* question, see Robinson et al. (1992)). MSQs can also exhibit more than one intent, presenting a challenge for both our classifier and the expert annotators (see Appendix E).

A limitation of our classifier is the focus on explicit MSQs, which can be identified with well-defined features. The low recall of our classifier indicates that MSQs are often implicit, missing certain markers or not completely fulfilling the distinguishing requirements. Figure 1b shows that while DISJUNCTIVE and CONDITIONAL MSQs are often explicitly signalled, the other types are likely to be implicit.

6 Conclusion

Inspired by the role of discourse relations in MSQ answering strategies, we propose a novel definition of five different categories of MSQs based on their corresponding speaker intents. We introduce a rich and diversified multi-sentence questions dataset, which contains 162,000 MSQs extracted from Stack Exchange. This achieves our goal of providing a resource for further study of MSQs. Additionally, we implement a baseline classifier based on surface features as a preliminary step towards successful answering strategies for MSQs.

Future work could improve on our classifier by considering implicit MSQs, with one potential approach being to transform explicit MSQs into implicit examples by removing some markers while ensuring the relation is still valid. Other areas for further work include implementing appropriate answering strategies for different types of MSQs, and investigating whether and how longer chains of MSQs differ compared to pairs of connected questions.

Acknowledgments

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh. We would like to thank Bonnie Webber for her supervision, and Ivan Titov and Adam Lopez for their useful advice.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2017. VQA: Visual question answering. *International Journal of Computer Vision*, 123(1):4–31, May.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Tudor Ciurca. 2019. Sense classification of multi-sentence questions. Master's thesis, School of Informatics, University of Edinburgh.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

H. P. Grice. 1975. Logic and conversation. In P Cole and J Morgan, editors, *Syntax and Semantics 3*, pages 41–58. Academic Press.

Jerry R Hobbs. 1979. Coherence and coreference. *Cognitive Science*, 3(1):67–90.

Zhiheng Huang, Marcus Thint, and Zengchang Qin. 2008. Question classification using head words and their hypernyms. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 927–936, Honolulu, Hawaii, October. Association for Computational Linguistics.

Blooma John and Jayan Kurian. 2011. Research issues in community based question answering. In *PACIS 2011 - 15th Pacific Asia Conference on Information Systems: Quality Research in Pacific*, page 29, 01.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *CoRR*, abs/1705.03551.

Andrew Kehler. 2006. Discourse coherence. In Laurence Horn and Gregory Ward, editors, *The Handbook of Pragmatics*, pages 241–265. Blackwell Publishing Ltd.

Vijay Krishnan, Sujatha Das, and Soumen Chakrabarti. 2005. Enhanced answer type inference from questions using sequential models. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 315–322, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea, July. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text: Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, March.

Jeffrey D Robinson, Nikolas Coupland, and Justine Coupland. 1992. "How Are You?": Negotiating Phatic Communion. *Language in Society*, 21(2):207–230.

Hannah Rohde, Anna Dickinson, Chris Clark, Annie Louis, and Bonnie Webber. 2015. Recovering discourse relations: Varying influence of discourse adverbials. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 22–31.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. *CoRR*, abs/1809.01494.

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: an assessment. *Natural Language Engineering*, 14(3):369–416.

Manfred Stede. 2011. Discourse processing. *Synthesis Lectures on Human Language Technologies*, 4(3):1–165, nov.

Akihiro Tamura, Hiroya Takamura, and Manabu Okumura. 2005. Classification of multiple-sentence questions. In *Second International Joint Conference on Natural Language Processing: Full Papers*.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn discourse treebank 3.0 annotation manual.

A Number of questions by topic

Topic	Number of questions	Topic	Number of questions
SuperUser	17197	Philosophy	799
ServerFault	14780	Hinduism	760
ElectricalEngineering	9847	PhysicalFitness	760
Arquade	9704	Homebrewing	671
Physics	8192	QuantFinance	667
English	7851	Outdoors	599
EnglishLearners	5922	Sports	576
SciFiFantasy	4920	Islam	524
InformationSecurity	4652	BiblicalHermeneutics	507
RPG	4259	Buddhism	492
DIY	3017	Engineering	492
Travel	2985	Linguistics	484
Academia	2488	TheorecticalCompSci	484
PersonalFinance	2312	Chess	483
StackOverflowMeta	2265	SoundDesign	482
SeasonedAdvice	2234	Pets	473
UX	2219	Economics	472
Workplace	2189	Parenting	472
Photography	2111	CognitiveSciences	444
Aviation	1949	Monero	425
Biology	1832	Health	420
Bitcoin	1806	ComputationalScience	398
Worldbuilding	1801	ReverseEngineering	392
MiYodeya	1769	EarthScience	386
Chemistry	1652	ProjectManagement	383
Music	1589	ArtificialIntelligence	360
ComputerScience	1500	Expatriates	310
Cryptography	1487	Robotics	295
GraphicDesign	1450	AmateurRadio	285
Ethereum	1415	Woodworking	283
BoardGames	1402	Literature	234
SpaceExploration	1354	HistoryScienceMathematics	229
Motors	1285	Tor	219
Bicycles	1241	Lego	213
Anime	1208	MartialArts	202
Law	1188	Mythology	174
NetworkEngineering	1160	InterpersonalSkills	156
Christianity	1084	Freelancing	148
Politics	1074	Poker	145
History	1066	Genealogy	137
Gardening	1007	MusicFans	129
DataScience	955	ArtsAndCrafts	118
SignalProcessing	912	Movies	118
Puzzling	888	SustainableLiving	109
Astronomy	816	OpenData	105
Skeptics	807	LifeHacks	83
Writers	805		

Table 2: Number of questions extracted by site topic in StackOverflow dataset, sorted by descending number of questions

B Contrastive features

Note that these requirements do not apply in all cases: if all conditions are met then we assert that a pair of questions *must* be of that type, but the absence of a feature does not *forbid* that relation from being present. A list of the lexical markers used to define features is given in Appendix C. Like discourse relations in declaratives, question discourse relations may be *implicit*, i.e. not marked with a connective or other marker but inferable to the listener. These implicit relations continue to be very challenging for automatic systems (Sporleder and Lascarides, 2008) and we do not attempt to handle them.

Feature	Surface	Separable	Reformulated	Disjunctive	Conditional	Elaborative	Elab. Statement
Anaphora	Pronoun in Q2	+	-			+	
	VP ellipsis in Q2	-					
Polarity	Polar Q1	-		+	+		
	Polar Q2			+			-
	Wh Q1			-	-		
	Wh Q2			-		-	-
	Q2 = Statement	-	-	-		-	+
Disc. Marker	"if" marker	-	-	-	+	-	-
	"or"	-	-	+	-	-	-
	"elab." marker	-	-	-	-	+	
	"sep." marker		-	-	-	-	-
Semantics	Word vector		+				+

Table 3: Definitions for each MSQ type. '+' indicates that a feature is *required*, while '-' means the feature is *disallowed*. Types can be ignored for some features, meaning that the features are neither disallowed nor required.

We include the case where Q2 is a statement, as in "*How can I transport my cats if I am moving a long distance? For example, to London.*". Although these forms of MSQ do not appear in our dataset due to the filtering method used, they are in general valid, and we include them for completeness.

To evaluate semantic similarity between Q1 and Q2, we calculated the cosine similarity between the mean of the words vectors, and compared to a threshold of 0.8.

C Lexical Markers

Some of the discourse markers are drawn from the Penn Discourse Tree Bank (PDTB) annotation scheme (Webber et al., 2019).

C.1 Anaphora Markers

To identify cases of anaphora, we searched for the following pronoun strings (and *it's*) in the second question:

she, he, it, they, her, his, its, their, them, it's

C.2 Verb Ellipsis Markers

To identify cases of verb ellipsis, we searched for the following pro-forms of full verb phrases in the second question:

do so, did so, does so, do it, do too, does too, did too, did it too, do it too, does it too

C.3 Polar Question Markers

do, does, did, didn't, will, won't, would, is, are, were, weren't, wasn't, can, can't, could, must, have, has, had, hasn't, haven't, should, shouldn't, may, might, shall, ought

C.4 Wh- words

who, what, where, when, why, how, which

C.5 Conditional (*"if"*) Markers

if so, accordingly, then, as a result, it follows, subsequently, consequently, if yes, if not, if the answer is yes, if the answer is no

C.6 Elaborative Markers

for instance, for example, e.g., specifically, particularly, in particular, more specifically, more precisely, therefore

C.7 Separable Markers

also, secondly, next, related, relatedly, similarly, furthermore

D Classifier visualisation

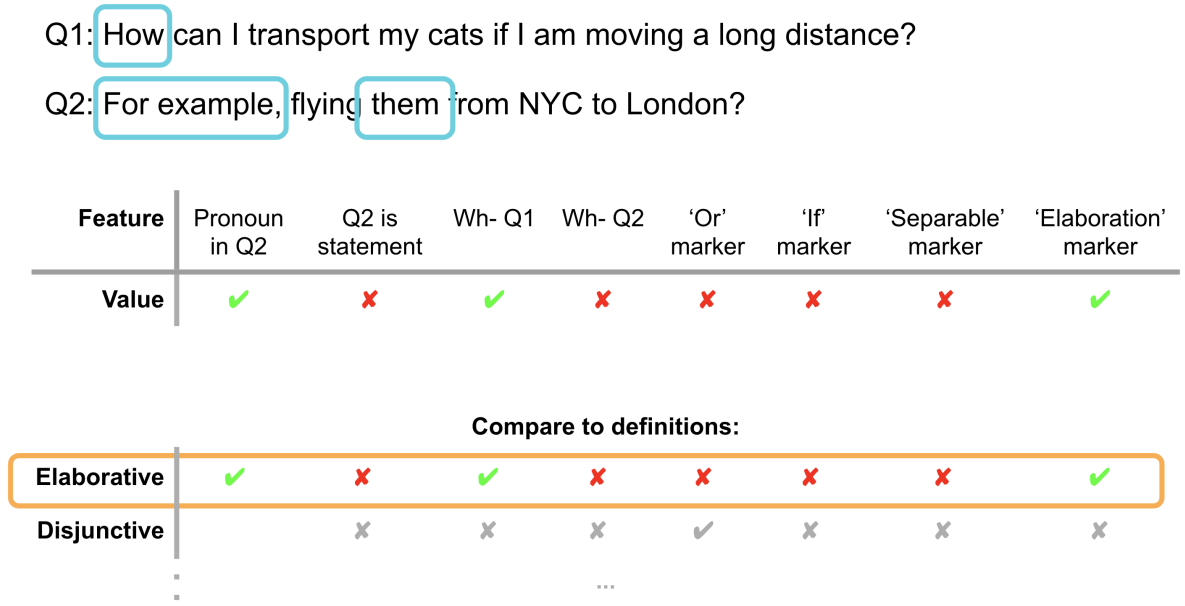


Figure 2: A visualisation of the labelling process. The system checks for the presence of each feature in the input text (circled in blue) and constructs a feature vector for the pair of questions. This feature vector is compared to the definitions, and if a match is found the questions are assigned the corresponding label. Note that not all features are shown.

E Annotator agreement

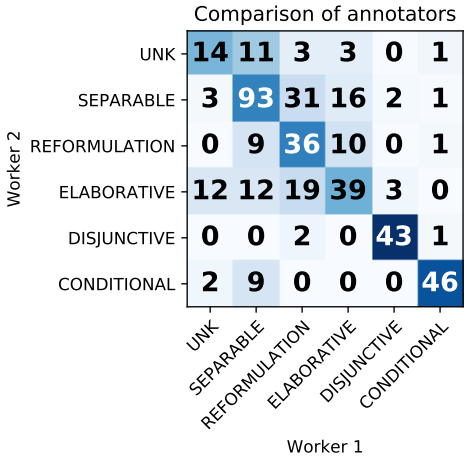


Figure 3: Confusion matrix between the two annotators who labelled our test set. While there is good agreement on the MSQ types that are often explicitly signalled (DISJUNCTIVE and CONDITIONAL), the other types are often more subtle, and examples may involve multiple intents.