Query-based sentence fusion is better defined and leads to more preferred results than generic sentence fusion*

Emiel Krahmer
Tilburg University
Tilburg, The Netherlands
E.J.Krahmer@uvt.nl

Erwin Marsi
Tilburg University
Tilburg, The Netherlands
E.C.Marsi@uvt.nl

Paul van Pelt
Tilburg University
Tilburg, The Netherlands
paul.vanpelt@gmail.com

Abstract

We show that question-based sentence fusion is a better defined task than generic sentence fusion (Q-based fusions are shorter, display less variety in length, yield more identical results and have higher normalized Rouge scores). Moreover, we show that in a QA setting, participants strongly prefer Q-based fusions over generic ones, and have a preference for union over intersection fusions.

1 Introduction

Sentence fusion is a text-to-text generation application, which given two related sentences, outputs a single sentence expressing the information shared by the two input sentences (Barzilay and McKeown 2005). Consider, for example, the following pair of sentences:[1]

(1) Posttraumatic stress disorder (PTSD) is a psychological disorder which is classified as an anxiety disorder in the DSM-IV.

(2) Posttraumatic stress disorder (abbrev. PTSD) is a psychological disorder caused by a mental trauma (also called psychotrauma) that can develop after exposure to a terrifying event.

Fusing these two sentences with the strategy described by Barzilay and McKeown (based on aligning and fusing the respective dependency trees) would result in a sentence like (3).

(3) Posttraumatic stress disorder (PTSD) is a psychological disorder.

Barzilay and McKeown (2005) argue convincingly that employing such a fusion strategy in a multi-document summarization system can result in more informative and more coherent summaries.

It should be noted, however, that there are multiple ways to fuse two sentences. Besides fusing the shared information present in both sentences, we can conceivably also fuse them such that *all* information present in either of the sentences is kept, without any redundancies. Marsi and Krahmer (2005) refer to this latter strategy as **union fusion** (as opposed to **intersection fusion**, as in (3)). A possible union fusion of (1) and (2) would be:

(4) Posttraumatic stress disorder (PTSD) is a psychological disorder, which is classified as an anxiety disorder in the DSM-IV, caused by a mental trauma (also called psychotrauma) that can develop after exposure to a terrifying event.

Marsi and Krahmer (2005) propose an algorithm which is capable of producing both fusion types. Which type is more useful is likely to depend on the kind of application and information needs of the user, but this is essentially still an open question.

*Thanks are due to Ed Hovy for discussions on the Rouge metrics and to Carel van Wijk for statistical advice. The dataset described in this paper (2200 fusions of pairs of sentences) is available upon request. This research was carried out within the Deaso project (http://daeso.uvt.nl/).

[1] All examples are English translations of Dutch originals.

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 193–196,
Columbus, Ohio, USA, June 2008.

However, there is a complication. Daumé III & Marcu (2004) argue that generic sentence fusion is an ill-defined task. They describe experimental data showing that when participants are given two consecutive sentences from a single document and are asked to fuse them (in the intersection sense), different participants produce very different fusions. Naturally, if human participants cannot reliably perform fusions, evaluating automatic fusion strategies is always going to be a shaky business. The question is *why* different participants come to different fusions. One possibility, which we explore in this paper, is that it is the generic nature of the fusion which causes problems. In particular, we hypothesize that fusing two sentences in the context of a preceding question (the natural setting in QA applications) results in more agreement among humans. A related question is of course what the results would be for union fusion. Will people agree more on the unions than on the intersections? And is the effect of a preceding question the same for both kinds of fusion? In Experiment I, below, we address these questions, by collecting and comparing four different fusions for various pairs of related sentences, both generic and question-based ones, and both intersection and union ones.

While it seems a reasonable hypothesis that question-based fusions will lead to more agreement among humans, the really interesting question is which fusion strategy (if any) is most appreciated by users in a task-based evaluation. Given that Experiment I gives us four different fusions per pair of sentence, an interesting follow-up question is which leads to the best answers in a QA setting. Do participants prefer concise (intersection) or complete (union) answers? And does it matter whether the fusion was question-based or not? In Experiment II, we address these questions via an evaluation experiment using a (simulated) medical question-answering system, in which participants have to rank four answers (resulting from generic and question-based intersection and union fusions) for different medical questions.

2 Experiment I: Data-collection

Method To collect pairs of related sentences to be fused under different conditions, we proceeded as follows. As our starting point we used a set of 100 medical questions compiled as a benchmark for evaluating medical QA systems, where all correct answers were manually retrieved from the available text material. Based on this set, we randomly selected 25 questions for which more than one answer could be found (otherwise there would be nothing to fuse), and where the first two answer sentences shared at least some information (otherwise intersection fusion would be impossible).

Fusion type	**Length M (SD)**	**# Id.**
Generic Intersection	15.6 (2.9)	73
Q-Based Intersection	8.1 (2.5)	189
Generic Union	31.2 (7.8)	109
Q-Based Union	19.2 (4.7)	134

Table 1: Mean sentence length (plus Standard Deviation) and number of identical fusion results as a function of fusion type ($n = 550$ for each type).

Participants were 44 native speakers of Dutch (20 women) with an average age of 30.1 years, none with a background in sentence fusion research. Experiment I has a mixed between-within subjects design. Participants were randomly assigned to either the intersection or the union condition, and within each condition they first had to produce 25 generic and then 25 question-based fusions. In the latter case, participants were given the original question used to retrieve the sentences to be fused.

The experiment was run using a web-based script. Participants were told that the purpose of the experiment was merely to gather data, they were not informed about our interest in generic vs question based fusion. Before participants could start with their task, the concept of sentence fusion (either fusion or intersection, depending on the condition) was explained, using a number of worked examples. After this, the actual experiment started.

Results First consider the descriptive statistics in Table 1. Naturally, intersection fusion leads to shorter sentences on average than union fusion. More interestingly, question (Q)-based fusions lead to significantly shorter sentences than their generic counterparts (intersection $t = 9.1, p < .001$, union: $t = 6.1, p < .001$, two-tailed). Also note that

	Generic Intersection	Q-Based Intersection	Generic Union	Q-Based Union
Rouge-1	.036	.068	.035	.041
Rouge-SU4	.014	.038	.018	.020
Rouge-SU9	.014	.040	.016	.020

Table 2: Average Rouge-1, Rouge-SU4 and Rouge-SU9 (normalized for sentence length) as a function of fusion type.

the variation among participants decreases in the Q-based conditions (lower standard deviations). This suggests that participants in the Q-based conditions indeed show less variety in their fusions than participants in the generic conditions. This is confirmed by the number of identical (i.e., duplicated) fusions, which is indeed higher in the Q-based conditions, although the difference is only significant for intersections ($\chi^2(1) = 51.3, p < .001$).

We also computed average Rouge-1, Rouge-SU4 and Rouge-SU9 scores for each set of fusions, to be able to quantify the overlap between participants in the various conditions. One complication is that these metrics are sensitive to sentence-length (longer sentences are more likely to contain overlapping words than shorter ones), hence in Table 2 we report on Rouge scores that are normalized with respect sentence length. The resulting picture is surprisingly consistent: Q-based fusion on all three metrics results in higher normalized Rouge scores, where the difference is generally small in the case of union, and rather substantial for intersection.

3 Experiment II: Evaluation

The previous experiment indicates that Q-based fusion is indeed a better-defined summarization task than generic fusion, in this experiment we address the question which kind of fusion participants prefer in a QA application.

Method We selected 20 from the 25 questions used in Experiment I, for which we made sure that the fusions in the four categories resulted in sentences with a sufficiently different content. For each question, one representative sentence was selected from the 22 fusions produced by participants in Experiment I, for each of the four categories (Q-based intersection, Q-based union, Generic intersection and Generic union). This representative sentence was the most frequent result for that particular category. When no such sentence was present for a particular task, a random selection was made.

Fusion type	Mean Rank
Q-Based Union	1.888
Q-Based Intersection	2.471
Generic Intersection	2.709
Generic Union	2.932

Table 4: Mean rank from 1 (= "best") to 4 (="worst") as a function of fusion type.

Participants were 38 native speakers of Dutch (17 men), with an average age of 39.4 years. None had participated in Experiment I and none had a background in sentence fusion research. Participants were confronted with the selected 20 questions, one at a time. For each question, participants saw four alternative answers (one from each category). Figure 3 shows one question, with four different fusions derived by participants from example sentences (1) and (2). Naturally, the labels for the 4 fusion strategies were not part of the experiment. Participants were asked to rank the 4 answers from "best" (rank 1) to "worst" (rank 4), via a forced choice paradigm (i.e., they also had to make a choice if they felt that two answers were roughly as good). Experiment II had a within-subjects design, which means that all 38 participants ranked the answers for all 20 questions.

Results Table 4 gives the mean rank for the four fusion types. To test for significance, we performed a repeated measures Analysis of Variance (ANOVA) with fusion type and question as the independent variables and average rank as the dependent variable. A main effect was found of fusion type ($F(3, 111) = 20.938, p < .001, \eta^2 = .361$).

	What is PTSD?
Generic Intersection	Posttraumatic stress disorder (PTSD) is a psychological disorder.
Q-based Intersection	PTSD stands for posttraumatic stress disorder and is a psychological disorder.
Generic Union	Posttraumatic stress disorder (PTSD) is a psychological disorder, which is classified as an anxiety disorder in the DSM-IV, caused by a mental trauma (also called psychotrauma) that can develop after exposure to a terrifying event.
Q-based Union	PTSD (posttraumatic stress disorder) is a psychological disorder caused by a mental trauma (also called psychotrauma) that can develop after exposure to a terrifying event.

Table 3: Example question from Experiment II, with four possible answers, based on different fusions strategies (obtained in Experiment I).

Pairwise comparisons using the Bonferroni method show that all comparisons are statistically significant (at $p < .001$) except for the one between Generic Intersection and Generic Union. Thus, in particular: Q-based union is ranked significantly higher than Q-based intersection, which in turn is ranked significantly higher than both Generic union and intersection (whose respective ranks are not significantly different).

The ANOVA analysis also revealed a significant interaction between question and type of fusion ($F(57, 2109) = 7.459, p < .001, \eta^2 = .168$).[2] What this means is that relative ranking varies for different questions. To better understand this interaction, we performed a series of Friedman tests for each question (the Friedman test is a standard non-parametric test for ranked data). The Friedman analyses revealed that the overall pattern (Q-based union $>$ Q-based intersection $>$ Generic Union / Intersection) was found to be significant for 13 out of the 20 questions. For four of the remaining seven questions, Q-based union ranked first as well, while for two questions Q-based intersection was ranked as the best answer. For the remaining question, there was no significant difference between the four fusion types.

4 Conclusion and discussion

In this paper we have addressed two questions. First: is Q-based fusion a better defined task than generic fusion? Here, the answer seems to be "yes": Q-based fusions are shorter, display less variety in length, result in more identically fused sentences and have higher normalized Rouge scores, where the differences are larger for intersection than for union. Inspection of the fused sentences reveals that there is simply more potential variation on the word level (do I select this word from one input sentence or from the other?) for union fusion than for intersection fusion. Second: which kind of fusion (if any) do users of a medical QA system prefer? Here a consistent preference order was found, with rank 1 = Q-based union, rank 2 = Q-based Intersection, rank 3/4 = Generic intersection / union. Thus: participants clearly prefer Q-based fusions, and prefer more complete answers over shorter ones.

In future research, we intend to collect new data with different questions per sentence pair, to find out to what extent the question and its phrasing drive the fusion process. In addition, we will also let sentences from different domains be fused, based on the hypothesis that fusion strategies may differ across domains.

[2] Naturally, there can be no main effect of question, since there is no variance; the ranks 1-4 are fixed for each question.

References

Regina Barzilay and Kathleen McKeown. 2005. Sentence Fusion for Multidocument News Summarization. *Computational Linguistics*, 31(3), 297-328.

Hal Daumé III and Daniel Marcu. 2004. Generic Sentence Fusion is an Ill-Defined Summarization Task. *Proceedings of the ACL Text Summarization Branches Out Workshop*, Barcelona, Spain.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using N-gram co-occurrence statistics. *Proceedings of NAACL '03*, Edmonton, Canada.

Erwin Marsi and Emiel Krahmer. 2005. Explorations in Sentence Fusion. *Proceedings of the 10th European Workshop on Natural Language Generation*, Aberdeen, UK.