Applications and Discovery of Granularity Structures in Natural Language Discourse

Rutu Mulkar-Mehta, Jerry Hobbs and **Eduard Hovy**
me@rutumulkar.com, hobbs@isi.edu, hovy@isi.edu
Information Sciences Institute
University of Southern California

Abstract

Granularity is the concept of breaking down an event into smaller parts or granules such that each individual granule plays a part in the higher level event. Humans can seamlessly shift their granularity perspectives while reading or understanding a text. To emulate such a mechanism, we describe a theory for inferring this information automatically from raw input text descriptions and some background knowledge to learn the global behavior of event descriptions from local behavior of components. We also elaborate on the importance of discovering granularity structures for solving NLP problems such as – automated question answering and text summarization.

1 Introduction

"Granularity" can be defined as a concept which involves breaking up a coarse and complex phenomenon into finer and simpler phenomena. This phenomenon can be anything from business processes to scientific processes to everyday activities. We use granularity concepts in our everyday life for the purposes of planning. For instance, consider the activity of driving to the grocery store. It involves some fine-grained events like opening the car door, starting the engine, planning the route and driving to the destination. Each of these could further be decomposed into finer levels of granularity. For instance, planning the route might involve entering an address into GPS and following directions. Granularity concepts are often reflected in natural language discourse. Newspaper articles are a classical example of granularity shifts in natural language discourse, with a coarse high level description in the first paragraph, and a more detailed fine granularity description in the subsequent paragraphs.

Granularity theories have been developed in various areas of research such as philosophy (Bittner and Smith 2001), theoretical computer science and ontology (Keet 2008) and natural language processing (Mani 1998; Hobbs 1985). However, none of them discuss granularity as it exists in natural language discourse or explored whether granularity structures can be identified and extracted from natural language. In our previous work (Mulkar-Mehta, Hobbs, and Hovy 2011), we describe a theory of granularity in natural language discourse and an annotation study to validate this theory. In this paper, we present the summary of our granularity theory (Section 2), and take this concept further by providing an outline of an algorithm that can be used to discover and extract granularity structures in natural language discourse (Section 3). We finally describe how extracting granularity structures in natural language texts can assist in solving NLP problems of question answering and text summarization (Section 4).

2 Theory of Granularity in Natural Language

Humans can easily shift through various levels of granularity for textual understanding. However, for automated granularity identification and extraction, it is important to explicitly recognize the identifiers that indicate a shift in granularity. We propose the following theory for modeling granularity in Natural Language Discourse.

A granularity structure exists only if at least 2 levels of information are present in text, such that the events in the coarse granularity can be decomposed into the events in the fine granularity and the events in the fine granularity combine together to form at least one segment of the event in the coarse granularity. Three types of relations exist between the objects in coarse and fine granularity: *part-whole relationship* between entities, *part-whole relationship between events*, and *causal relationship* between the fine and coarse granularity. These relations signal a shift in granularity. A graphical representation of our theory of granularity in natural language is shown in Figure 1.

In Figure 1, G_c represents the phrase or sentence with coarse granularity information and G_f represents a phrase or sentence with fine granularity information. Three possible links connect the objects of coarse granularity and the objects of fine granularity - *part-whole relations* between events, *part-whole relations* between entities, and a *causal relation* between the events in the fine granularity and the events in the coarse granularity. The coarse granularity description gives us a high level overview of an event,

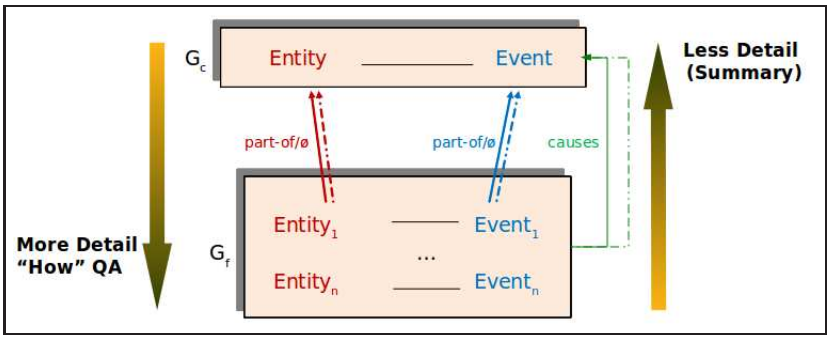

Figure 1: Granularity in Natural Language Descriptions

abstracting away the details and looking at the big picture. The fine granularity description provides the details of how the event happened, abstracting away the big picture of the overall effect of the fine grained events. At least one of these relations must be present to infer a granularity structure in the text, in which case the other relations may be inferred (represented by the dotted lines in Figure 1). Instantiating text phrases into this model produces granularities of text. Consider the following simplified paragraph extracted from a newspaper article describing a football game:

San Francisco 49ers lost 27-17 to the Green Bay Packers. Brett Favre threw a three-yard touchdown pass to Keith Jackson in the first quarter moving the 49ers ahead 7-0. Brett Favre threw a 13-yard touchdown toss to Mark Chmura in the second quarter moving the 49ers ahead 14-0.

Figure 2 shows the instantiation of the paragraph into three levels of granularity, with the top level representing the coarsest granularity, and the bottom level representing the finest granularity. The part-whole links between entities (*Brett Favre is part of San Francisco 49ers*) and events (*touchdown is part of a quarter, quarter is part of a game*) are marked. Causality is indicated by the word *move*, linking two levels of granularity. The presence of *causal* and *part-whole* relations indicates a shift in granularity in the sentence, and the individual events in the sentence can be split into two levels of granularity. The dotted line represents an inferred causal relationship because of the absence of explicit causal markers.

The evaluation of our causal granularity theory is provided in a separate compilation (Mulkar-Mehta, Hobbs, and Hovy 2011), where we statistically prove this feature set using a human annotation study for granularity identification. We achieve an average pairwise kappa (Cohen 1960) agreement score of 0.85.

3 Proposed Pipeline for Automatic Discovery of Causal Granularity Structures

This section describes the basic building blocks and proposed algorithm for automatic discovery of causal granularity structures from discourse. There are four components in the granular causality extraction pipeline: Identification of smallest discourse segment for analysis (Section 3.1); Creating a background knowledge base of Part-Whole relations (Section 3.2); Inferencing Causal Granularity (Section 3.3); Evaluation of the final results (Section 3.4). Algorithm 1 summarizes these sections and is the proposed algorithm for extraction of granularity structures from text.

Algorithm 1 Algorithm for Automatic Discovery of Causal Granularity Structures

1: Obtain part of relations (Pev_1,Wev_1...Pev_n,Wev_n) and (Pen_1,Wen_1...Pen_n,Wen_n)
2: **for all** Article A_n **do**
3: Obtain sentences (S_1...S_m) in A_n
4: **end for**
5: **for all** S_i,S_j in A_a **do**
6: **for all** (Pev_k,Wev_k), $k = 1$ to n **do**
7: **if** $Pev_k \in S_i$ and $Wev_k \in S_j$ **then**
8: **for all** (Pen_q,Wen_q), $q = 1$ to m **do**
9: **if** $Pen_q \in S_i$ and $Wen_q \in S_j$ **then**
10: **Inference:** S_i **causes** S_j
11: **end if**
12: **end for**
13: **end if**
14: **end for**
15: **end for**
16: Evaluate the Causal Granularity Relations using Annotations

3.1 Identification of Smallest Discourse Segment for Analysis

In order to identify shifts in granularity in discourse, we first need to identify the discourse unit which can represent a single level of granularity. A granularity level can often span multiple sentences, as shown in our previous work (Mulkar-Mehta, Hobbs, and Hovy 2011), but for simplification purposes, we will consider a single sentence as a smallest discourse segment for analysis. In a given article (A_a), the discourse segments of analysis are S_i...S_n, where the article has n sentences.

Figure 2: Example of Granularity in Natural Language Text

3.2 Automatic Part-Whole relation extraction

There are 2 types of meronymic part-of relations required for causal granularity extraction - *entity part-whole relations* (Pen,Wen) and *event based part-whole relations* (Pev,Wev).

(Winston, Chaffin, and Herrmann 1987) discuss meronymic relations and a taxonomy for representing them. They introduce six types of part-whole relationships: (i) Component-Integral (e.g., *pedal* is a component of the integral *bike*), (ii) Member-Collection (e.g., a *ship* is a member of the collection, a *fleet*), (ii) Portion-Mass (e.g., a *slice* is a portion of the mass, a *pie*), (iv) Stuff-Object (e.g., *steel* is one of the ingredients/stuff of the object *car*), (v) Feature-Activity (e.g., *paying* is one of the features of the whole activity of *shopping*), (vi) Place-Area (e.g., *Everglades* is a place within the area of *Florida*). For discovery of granularity relations, the Feature-Activity type relation is the event based part-whole relation and the remaining categories are part-whole relations for entities.

Several initiatives such as (Girju, Badulescu, and Moldovan 2003), (Hage, Kolb, and Schreiber 2006) and (Ittoo et al. 2010) have attempted to extract general part-whole relations from discourse. As a first step, we will re implement these state of the art techniques for part-whole relation extraction to obtain part of relations for events (Pev_1,Wev_1...Pev_n,Wev_n) and part of relations for entities (Pen_1,Wen_1...Pen_n,Wen_n).

3.3 Inferencing Causal Granularity

In this step we consider two sentences in the corpus S_1 and S_2 and check whether there exists (Pev,Wev) and (Pen,Wen) pairs in the sentence pair, where Pev and Pen lie in S_1 and Wev and Wen lie in S_2. We can then derive the inference that S_1 causes S_2, where S_1 contains the event and entity parts and S_2 contains the event and entity wholes.

3.4 Evaluation

Evaluation measures of the inferred relations will be obtained by passing the inferred relations to Mechanical Turk and using crowd sourcing to judge the precision of the inferred relations in the sentence pair. A gold standard will be developed using the Mechanical Turk annotations.

4 Applications of Granularity Structures for solving NLP problems

Having described the theory of granularity in text, and the algorithm for extracting such relations, we now describe different areas of NLP which will benefit from granularity relations. This section focuses is on 2 areas – Automated Question Answering and Text Summarization.

4.1 Automated Question Answering

Question answering has achieved high degrees of precision and recall for fact based questions such as - what, when, where (Hovy et al. 2001). However, this field has few contributions for answering causality based questions. “How” and “why” are two ways of asking causal questions in English. Both of these question types remain the most difficult types of questions to answer by automated question answering techniques. New methods and models have been introduced such as (Mrozinski, Whittaker, and Furui 2008; Verberne 2009) for answering these style of questions. However, it has been largely overlooked that cause and effect relations might not always be sequential, but might happen at the same time, where an event happening to a part entity causes an event to a whole entity. For example, *a car stops working when the engine breaks down, a team wins a game when an individual scores, a building collapses when the roof caves in*, and so on. In these cases, there might not be a sequential causality present between these events, but a granular causality. A very effective way to extract such relations is using granularity structures. This can be achieved in the following manner.

The theory of granularity is shown in Figure 1, repre-

senting two levels of granularity and three types of relations between them: *part-whole relations between events*, *part-whole relations between entities* and *causal connectives*. For instance, consider Figure 2 that instantiates the theory of granularity to the paragraph mentioned in Section 2.

Asking a question about the coarse granularity from Figure 2, one could ask the question: *How did the San Francisco 49ers win the game?*, to which the answer can be provided by going down one level of granularity and answering: *because they went ahead 7-0 in the first quarter and 14-0 in the second quarter.* Another possible question is: *How did the San Francisco 49ers move ahead in the first quarter?*, to which the reply would be from a finer granularity level : *Brett Favre threw a three-yard touchdown pass*.

The first step to achieving such a QA system is to identify different granularities of information from a given text and instantiating the granularity model with the texts. We already have an evaluation study of this work, and will present the results as a separate compilation (Mulkar-Mehta et al. in review).

4.2 Text Summarization

The granularity model represents a single time slice and all the events happening within it belonging to different grain sizes. If an entire text of information is represented into this granularity model, getting a summary of the events involves going up in the model to find the coarser grained information. For example, consider a game of football, where a lot of individual scoring events of touchdowns, field goals or the kick after a touchdown are present. If one asks for the summary of the first quarter, one can look at all the events happened in the first quarter and present the final results. Similarly, if one is interested in the final outcome of the game, the system can abstract all the fine grained information and prove just the top level information in the coarsest granularity.

If such a multi-granular structure is obtained from text, a text summary can be obtained by going up the granular hierarchy, abstracting away the low level details from the text. For instance, consider Figure 2. If the summary of the first quarter is required, one can give the coarsest granularity information describing the first quarter which is *the 49ers moved ahead 7-0*. Similarly, if one wants the summary of the game, we could go to the coarsest granularity level describing the game and answer *San Francisco 49ers won the game 27-14.*

5 Conclusion

In this paper we present the theory of granularity for natural language texts and an algorithm for extraction of such granularity structures from discourse. Finally we present different applications of the theory of granularity for solving NLP problems of automated question answering and text summarization.

As a part of the future work we have already started working on developing a system for automatic granularity extraction. We will compare this with the state of the art techniques for answering causality style questions to empirically evaluate the significance of granularity structures for question answering. Results of our system for automatic extraction of granularity can be found in a separate compilation (Mulkar-Mehta et al. in review).

References

Bittner, T., and Smith, B. 2001. Granular partitions and vagueness. *Proceedings of the international conference on Formal Ontology in Information Systems - FOIS '01*.

Cohen, J. 1960. A coefficientof agreement for nominal scales. *Educational and Psychological Measurement* 20.

Girju, R.; Badulescu, A.; and Moldovan, D. 2003. Learning semantic constraints for the automatic discovery of part-whole relations. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03* 1:1–8.

Hage, W. R. V.; Kolb, H.; and Schreiber, G. 2006. A Method for Learning Part-Whole Relations. *Proceedings of the 5th International Semantic Web Conference (ISWC 2006)* 4273:723 – 736.

Hobbs, J. R. 1985. Granularity. *In Proceedings of the Ninth International Joint Conference on Artificial Intelligence* 432–435.

Hovy, E.; Gerber, L.; Hermjakob, U.; Lin, C.-Y.; and Ravichandran, D. 2001. Toward semantics-based answer pinpointing. *Proceedings of the first international conference on Human language technology research - HLT '01*.

Ittoo, A.; Bouma, G.; Maruster, L.; and Wortmann, H. 2010. Extracting Meronymy Relationships from Domain- Specific , Textual Corporate Databases. *NLDB* 48–59.

Keet, C. M. 2008. *A Formal Theory of Granularity*. Ph.D. Dissertation, Faculty of Computer Science, Free University of Bozen-Balzano, Italy, Oxford, UK.

Mani, I. 1998. A Theory of Granularity and its Application to Problems of Polysemy and Underspecification of Meaning. *In Principles of Knowledge Representation and Reasoning: Proceedings of the Sixth International Conference (KR'98)* 245–255.

Mrozinski, J.; Whittaker, E.; and Furui, S. 2008. Collecting a Why-question corpus for development and evaluation of an automatic QA-system. *Association of Computational Linguistics* (June):443–451.

Mulkar-Mehta, R.; Hobbs, J. R.; and Hovy, E. 2011. Granularity in Natural Language Discourse. *International Conference on Computational Semantics* 360—-364.

Verberne, S. 2009. *In Search of the Why*. Ph.D. Dissertation, University of Nijmegen, Oxford, UK.

Winston, M. E.; Chaffin, R.; and Herrmann, D. 1987. A Taxonomy of Part-Whole Relations. *Cognitive Science* 11(4):417–444.