Annotating Topic Development in Information Seeking Queries

Marta Andersson*, Adnan Öztürel†, Silvia Pareti†
*English Language Department, Stockholm University, Sweden
†Google Inc., Brandschenkestrasse 110, 8002 Zürich, Switzerland
marta.andersson@english.su.se, {ozturel, spareti}@google.com

Abstract

This paper contributes to the limited body of empirical research in the domain of discourse structure of information seeking queries. We describe the development of an annotation schema for coding topic development in information seeking queries and the initial observations from a pilot sample of query sessions. The main idea that we explore is the relationship between constant and variable discourse entities and their role in tracking changes in the topic progression. We argue that the topicalized entities remain stable across development of the discourse and can be identified by a simple mechanism where anaphora resolution is a precursor. We also claim that a corpus annotated in this framework can be used as training data for dialogue management and computational semantics systems.

Keywords: corpus annotation, information structure, information seeking conversational query sessions

1. Introduction

Application of NLP techniques on the domain of information seeking queries is well worth exploring. The conversational setting between the query issuer who is seeking information and the dialogue management systems delivering it has a unique discourse structure. Deliverables of research on this specific discourse structure can be valuable in improving dialogue management systems. However, there is still very limited research in the literature on this topic mostly because of the lack of available data.

This paper contributes to a rather scarce body of empirical data on information-seeking queries (henceforth ISQ). The main goal pursued here is to devise an annotation methodology that can capture the discourse structure in a set of successive queries where each is information seeking in structure. We believe the methodology we present can serve well for preparing linguistic resources that can be used for training computational semantic applications, such as topic detection systems.

The aim of the present study is twofold: (i) to investigate the nature of topic development in discourse in a corpus of information-seeking queries, (ii) to identify the features that crucially participate in topic development in order to describe their role in this process.

By "*information seeking query sessions*" we mean a set of information-seeking queries (the issuer's input), where coherent discourse relations between the successive queries can be identified. We argue that in case where no referential ambiguity is present in the context of an information seeking query sessions, the progression of discourse topic can be identified (and also annotated) with a set of simple heuristic rules. However, in the case of referential ambiguity, which may be introduced by anaphora in follow-up queries, disambiguation can be achieved through automated anaphora resolution.

Recent advancements in computational semantics deliver methodologies to build wide-coverage systems that can construct meaning representations and carry out robust anaphora resolution (Bos, 2008). We believe our annotation methodology can set the ground for crafting linguistic resources that can be used to train specialized computational semantics systems.

In what follows, we present a consistent approach to strategies of topic continuation and shift that query issuers deploy in ISQ sessions. We have devised an experimental multi-layer annotation schema that can be used to capture, describe and evaluate phenomena related to topicality development in discourse. We have manually annotated a pilot sample of 200 English query sessions, where each session contained two successive information seeking questions.

Our annotation layers cover syntactic cues, semantic relations, discourse entities and discourse topic development. However, in this paper we only present the method to annotate discourse entities and topic development, and leave out syntactic and semantic categories, which have been commonly discussed in the literature. Throughout the annotation process, we identified several rules according to which discourse entity types can be identified and their roles mapped onto different types of discourse topic development.

2. Background

In the current study we adopted a modification of the linguistic notions of *Topic*, *Focus* and *discourse entity* that are used in information structure studies. In what follows we will therefore provide a brief discussion of the linguistic views on information packaging and specify how they differ from our approach.

The Topic-Focus distinction has been modeled in terms of presupposition (Strawson, 1964), referentiality and definite descriptions (Heim, 1982), hearer old/new information (Prince, 1992) and activation (Chafe, 1994; Lambrecht, 1996). Most commonly, this distinction assumes that Topic is the referent that the sentence is about, whereas Focus is what increases our knowledge about the Topic (Lambrecht, 1996). Consequently, Topic can be characterized as given, whereas Focus as new discourse information.

However, what we identify as Topic or Focus changes whether we consider **referential givenness/newness** (e.g.

existential presupposition, specificity, definiteness etc.) or **relational givenness/newness** (e.g. presupposition-focus, topic-comment etc.) (Gundel and Fretheim, 2004). The difference between referential and relational givenness/newness is that the former is not a linguistic concept, but relates to the states of knowledge in the speaker/hearer's mind. By contrast, relational givenness/newness is associated with the meanings and interpretations of the linguistic discourse representation and can be contextually determined.

From the annotation perspective it may be hard to consistently identify Topic and Focus based on referential givenness, where several different possibilities for interpretation of the same query exist and can potentially spawn disagreement between the annotators:[1]

(2.1) **A:** Did you order the chicken or the ***pork***?
B: It was the ***pork*** that I ordered.

In (2.1)-A, the "pork" is referentially given and therefore could be regarded as the Topic of (2.1)-B. However, it is also new in relation to the context of (2.1)-B and simply instantiates a variable component of the relationally given, topical part of the sentence, which is what the participant ordered. The expression thus yields new information and as a result can be considered a prominent focal element in this context.

If we were to identify what is expected by the answer, the referential approach would be useful, since the Focus would mark the element we need an answer for. However, for the present study, whose aim is to explore the topic development across pairs of queries, the relational approach is the most relevant. The focal point of interest of the present analysis is the topical relations between the content of the queries but not necessarily what is introduced in discourse by the anticipated answer. The relational approach to given/new enables us to identify the topic development in a straightforward way. The clear interconnection with the familiar linguistic phenomena makes this approach both useful and reliable for capturing information salient for the identification and interpretation of topic development type in the ISQs.

In this view anaphora resolution has an impact on the identification of topic progression; however, anaphora resolution is not the scope of the present study. The main idea pursued here is that the topicalized 'old' part of discourse is the information that can be retrieved not only via a grammaticalized referential expression such as pronoun or demonstrative, but also via omission, i.e. ellipsis and zero anaphora. Consequently, the element that cannot be omitted or anaphorically recalled is regarded as the focus of attention and hence Focus of the question. Consider a constructed example of an ISQ:

(2.2) $\mathbf{Q}_1$**:** When was ***Stockholm*** founded?
$\mathbf{Q}_2$**:** When was ***Zurich*** founded?

On the referential approach (which to a great extent sustains the mainstream linguistic view), the 'given' topical entities are Stockholm and Zurich (what the sentences are about), whereas the remaining context of the question is Focus, which asks about specifics related to those referents. This information is not considered given or activated yet, because it pertains to the content of the upcoming answer. Importantly, this content cannot be elided:

(2.3) $\mathbf{Q}_1$**:** When was Stockholm founded?
$\mathbf{Q}_2$**:** And Zurich?

Our analysis is concerned with the query issuer's informational needs (intention and purpose), which are likely to have prompted a given ISQ session. From this point of view the 'presupposed' (relationally given) information is the Topic of founding two different cities, which is the query part that can be replaced by another expression, omitted or elided (e.g. ISQ session in (2.3)). The focus of attention are the cities, the content that cannot be anaphorically retrieved. We believe that the topical relationship between the consecutive queries can be distinguished in this cohesive manner, where referential givenness based on traditionally understood anaphora resolution may but does not have to be the determinant of the issuer's needs. A more detailed discussion of this approach follows in section 3.1 below.

Few studies have addressed discourse structure of question answering interactions. The closest to our approach was proposed by Chai and Jin (2004). As they argue, questions carry distinctive discourse roles with respect to the whole discourse, which can be characterized in terms of informational content of the query. On this approach 'Content' has three major components, which are Target, Topic and Focus. Target indicates the expected answer type such as a proposition (e.g. for 'why' and 'how' questions), or a specific type of entity (e.g. 'time' and 'place'). Topic relates to the 'aboutness' or the scope of a question, whereas Focus indicates the current focus of attention given a particular topic and refers to a particular aspect of this topic.

The mainstay of Chai and Jin (2004) proposal is that the informational perspective of discourse should capture the semantics of the conveyed information. Consequently, Topic and Focus are linked with the semantic roles of the constituents in the question in terms of its predicate-argument structure (Gildea and Jurafsky, 2002). What follows is that Topic in this approach does not fully coincide with the linguistic definition, which usually involves an anaphorically retrievable Participant. Topic can be different discourse facets with different semantic roles - both participants (e.g. Agent) and activities, as in our example (2.2) above.

We follow Chai and Jin (2004)'s view on how Topic and Focus can be identified in ISQs and, consequently, how the topic progression can be tracked down and characterized. We also follow the topic development types proposed in their paper; however, the original model suffers from a lack of precise descriptions of how the abstract discourse roles can be pinned-down in the text and operationalized in order to identify topic development types. We provide a simpler and more systematically structured model. We distinguish between three types of discourse constructs, which we call 'discourse entities'.

Our notion of 'discourse entity' includes three abstract roles: Participant, Predicate and Property. This in con-

[1] Example 2.1 excerpted from (Gundel and Fretheim, 2004).

trast with Chai and Jin (2004), who discuss a semantic-rich model that categorizes Participant by semantic role (e.g. Agent), semantic type (e.g. human being) or id (e.g. Bill Clinton). In our approach Participant maps to the NP referent. We regard these distinctions as sufficient for the purpose of our study, for they can be efficiently linked to the discourse progression through Topic and Focus. The details and rationale behind this idea are described in 3.2 below.

3. Annotation Framework

The following sections provide the details of our annotation method including several corpus examples illustrating the steps and decisions taken in the process of corpus coding.

3.1. Current Approach to Topic and Focus

Defining Topic and Focus is the initial step towards devising a concrete annotation framework for discourse progression in ISQs. As mentioned, our approach does not fully correspond to the traditionally understood linguistic notions of Topic and Focus, but is akin to these concepts. The linguistic take on this distinction can be more formally described as follows:

- **Approach 1:** The Focus is what is elicited in the query and expected from the answer. It is a hint as to what the answer should contain. In this approach Topic is the given element and Focus is the new element.[2]

(3.1) **Q-A$_1$:** Age$_F$ of ***Obama***$_T$
Q-A$_2$: Age$_F$ of ***Clinton***$_T$
Q-B$_1$: Age$_F$ of ***Obama***$_T$
Q-B$_2$: Age$_F$ of ***Obama***$_T$

The approach we propose is based on a slightly different assumption:

- **Approach 2:** The Focus is the new element in the discourse that cannot be omitted (no anaphora on it). Topic is treated as the static element and Focus as the variable one.

(3.2) **Q-A$_1$:** ***Age***$_T$ of Obama$_F$
Q-A$_2$: ***Age***$_T$ of Clinton$_F$
Q-B$_1$: Age$_F$ of ***Obama***$_T$
Q-B$_2$: Size$_F$ of ***Obama***$_T$

While favoring sheer theoretical consistency, linguistically-oriented approach (Approach 1) is preferable, since the Topic simply conveys existential presupposition and Focus seeks information that is relevant and increases the knowledge about this Topic. This means that the new information obtained in the upcoming answer is the Focus of the question. However, as mentioned, the main goal of our analysis is to estimate the theme that the query issuer is interested in, based on the sole context of the ISQs. Therefore, what we attempt to identify are the fixed and variable components of the query context, which we find more informative about the issuer's goals than the information that can be retrieved via sole pronominalization.

[2]Topic denoted with *T* and Focus with *F* subscripts in the examples all through.

Specifically, we focus on the discourse entities that remain constant across discourse transitions, which we consider the Topic of the issuer's interest. This is illustrated in Q-A$_1$ and Q-A$_2$ query pair of Approach 2. In this query session the queries ask about a Property of two different Participants. In both queries the Property that is being asked about does not change. Therefore, we tag the constant discourse entity (Property *age*) as the Topic, whereas the altering discourse entites (Participants *Obama* and *Clinton*) are tagged as the Focus.

Distinctively, in follow up Q-B$_1$ and Q-B$_2$, where the Topic of the issuer's interest stays the same (Participant *Obama*) and the variable Focus is different Properties related to this Participant (*age* and *size*). In case of that particular query session the relations between discourse entities can be established via commonplace pronominalisation.

By contrast, a diagnostic test that we propose for topic identification in cases like Q-A$_1$ and Q-A$_2$ is the substitution of the topicalized entity with the phrase 'what-about'. This phrase signals the retention of the current Topic and in this respect resembles other referential expressions indicating the Topic - high accessibility markers consisting of less linguistic material (Ariel, 2001).[3]

By comparing both approaches we hypothesize that Approach 2 can result in a finer distribution of the annotated categories, whereas Approach 1 would cluster and condense the annotated distributions and cannot distinguish most of the phenomena where the query issuer explores various properties of a certain Topic. In parallel, next section elaborates the notion of 'discourse entities', which are the crucial components of our approach in annotating topic development in ISQs to study our hypothesis.

3.2. Discourse Entities

This section briefly presents the methodology to identify the discourse entities that are distinguished in this study. This level of analysis involves the division between the conceptual parts of a query into three interdependent tags: Participant, Predicate and Property. Unlike topic development types, all entities are tagged in each query in isolation. Once the entity types that are present in discourse and their roles are identified, the topic development type between the queries can be easily established as a follow up.

3.2.1. Participants

The category labeled *Participant* corresponds to nominal elements (common nouns referring to both animate and inanimate entities, proper names such as names of people, places, events e.g. *Christmas*, time periods e.g. *December*, measures etc. and also pronominals) in the ISQs. Thus, the

[3]We are aware that this test may not be operative in every context and is likely to depend on the verb arguments. Further, it should be noted that in cases where the anaphoric reference in Q$_2$ points back to the same Participant in Q$_1$, the 'what about' phrase does not exhibit the same anaphoric characteristics, for instance:
Q$_1$: "How old is Obama?"
Q$_2$: "What about his weight?"
This query pair would be an instance of topic extension participant – TEP (see section 3.3.2), where Obama is the topic of Q$_1$ and his weight the topic of Q$_2$. We treat such examples as special cases of discourse topic development.

Development Type	Description	Example
Topic Exploration (TEL)	The same Topic. Focus on a related Property.	Q_1: Who is Lady Gaga? Q_2: How old is she?
Topic Extension Participant (TEP)	The same Topic. Focus on another Participant.	Q_1: When did Lady Gaga start her career? Q_2: When did Madonna start her career?
Topic Extension Circumstances (TEC)	The same topic. Focus on time, place, etc.	Q_1: What's the time in New York? Q_2: What's the time in London?
Topic Extension Activity (TEA)	The same topicalized Participant Focus on different Predicates.	Q_1: When did Lady Gaga start her career? Q_2: When did she release "Poker Face"?
Topic Shift Activity (TSA)	Topic shifts from one Predicate to another related Predicate with different Participants.	Q_1: When did Lady Gaga play in Berlin? Q_2: How many people came to the concert?
Topic Shift Participant (TSP)	Topic shift from Predicate to a related Participant.	Q_1: When did Lady Gaga release "Poker Face"? Q_2: How long is this song?

Table 1: Topic development type tag set.

participant in (3.3)-Q_1 is ***Lady Gaga***, and in the follow-up query it is the pronoun ***she***.

(3.3) **Q_1:** How old is ***Lady Gaga***?
Q_2: When did ***she*** start singing?

Participants can be agentive (APTP) or non-agentive (NAPTP). An agentive participant is the one that undertakes an action. This fine-grained category includes all animate entities (e.g. humans and animals) as well as instances of figurative agency (e.g. "the computer won't cooperate"). In the follow-up query concerning Lady Gaga above, the artist is an agentive participant, because starting something is an agentive activity. A non-agentive participant, by contrast, is an experiencer of a state (e.g. he sleeps) or someone who takes part in a non-action event (e.g. he fell) (Jackendoff and Culicover, 2003).

3.2.2. Predicates

The category *Predicate* includes both main and auxiliary verbs and consists of three subcategories:

(a) **Events:** (i) action predicates (APRED e.g. *he went*), and (ii) non-action predicates (NPRED e.g. *he fell*).

(b) **Stative Predicates:** such as *be*, *sleep*, *love* (SPRED).

(c) **Procedural Predicates:** 'how to' or other types of 'how' queries (PPRED e.g. "*How do you make a lemon pie?*").

3.2.3. Properties

The category *Properties* (PROP) includes scales, comparison and measures, for instance:

(3.4) **Q_1:** How ***much is*** a British passport?
Q_2: How ***much is*** an Irish passport?

Certain states can also be categorized as Properties, as it is sometimes very difficult to distinguish between these two. This concerns adjectival passive voice constructions:

(3.5) **Q_1:** Who is Chris Pratt ***married to***?
Q_2: Who is she?

Being married is, admittedly, different from both the canonical property (e.g. *old*) and the canonical state (e.g. *asleep*). For this reason we propose to label such instances as 'event-like' property, which should help convey their ambiguity. Finally, most senses of the verb 'have' are also subsumed under this category.

3.3. Topic Development Types

The present section provides examples and brief descriptions of the topic development types. We identify Topic under each category based on accessibility and retrievability of discourse entities. Recall that the main rule we follow is that Focus is always the variable component of the query, which is new in relation to those entities (i.e. it is newly asserted or newly asked about), while Topic is the entity that remains constant. The topic development types we distinguish are summarized in Table 1.

3.3.1. Topic Exploration/Elaboration (TEL)

Topic of the queries remains the same, whereas the Focus explores its other aspect/peripheral (e.g. attributes, process, etc.). In (3.6) below, the topicalized Participant of Q_1 is anaphorically picked up as the topic of Q_2 and a request for additional information about this participant is made:

(3.6) **Q_1:** What do ***snails***$_T$ eat$_F$?
Q_2: How long can ***they***$_T$ be$_F$?

A common type of TEL are queries asking about Properties of the involved participants.

3.3.2. Topic Extension

The Topic remains the same in both queries, but the Focus involves new constraints such as time, location and participants, for instance:

(3.7) **Q_1:** What ***do*** snails$_F$ ***eat***$_T$?
Q_2: What ***do*** guinea pigs$_F$ ***eat***$_T$?

In these queries, query issuer ask about the same Predicate (eating) but there is a change of Participants (*snails* and *guinea pigs*). We refer to this type of discourse development as Topic Extension to Participant (TEP). This example is in line with our idea of Topic as a constant component of the query. In (3.7) this component is eating habbits of two animal species, unlike in the linguistic view, where that part of the queries would be considered Focus eliciting new information about the involved participants. In our approach, the snails and the guinea pigs convey new information (newly asked-about) in the contexts of the queries. These entities cannot be omitted in the consecutive queries, unlike the remaining context in which 'what-about' anaphora is operative.

Based on our corpus observations, we also suggest that Properties can make topicalized entities in exactly the same way, since their Participants instantiate relationally new information in the queries:

(3.8) **Q_1:** How ***old is***$_T$ Bill Clinton$_F$?
Q_2: How ***old is***$_T$ Barack Obama$_F$?

Another category of Topic extension is constraint change (TEC). For the time being we follow Chai and Jin (2004) and distinguish between two types of constraints; temporal or spatial:

(3.9) **Q_1:** What ***is there to do***$_T$ in Cocoa Beach Florida$_F$?
Q_2: What ***is there to do***$_T$ in Titusville Florida$_F$?

Another special case of Topic extension that we propose distinctive from the literature is Topic Extension to Activity (TEA) and stems from our corpus observations:

(3.10) **Q_1:** When was ***Peter Blake***$_T$ born$_F$?
Q_2: Where did ***he***$_T$ study art$_F$?

Both queries ask about the same Participant, but 'what-about' test does not apply in this case. However, the Topic entity is anaphorically retrieved from the previous context and retained as the doer behind the action conveyed (Peter Blake could be retrieved anaphorically and so this entity belongs to the group of non-variable discourse components).

3.3.3. Topic Shift

This topic progression type does not involve changes that qualify two queries as unrelated, such as asking about unrelated discourse entities. It might involve more subtle changes, such as:

(3.11) **Q_1:** Who is ***Abraham***$_T$ in the Bible$_F$?
Q_2: Who ***wrote***$_T$ the Old Testament$_F$?

The Topic of (3.11)-Q_1 is the identity of certain discourse Participant (Abraham) and hence this participant, whereas in (3.11)-Q_2 the Topic shifts to the activity of writing the Old Testament, which is the peripheral information. This shift is labeled Topic Shift to Activity (TSA). In a similar way, the Topic of the queries can also shift between the activity and participant (TSP):

(3.12) **Q_1:** When ***did*** Klimt ***paint***$_T$ Adele Bloch-Bauer$_F$?
Q_2: How much was ***it***$_T$ worth at the auction in New York$_F$?

4. Corpus

In order to test and develop the annotation framework described in this paper, we collected and annotated two small pilot corpora of information seeking query sessions. The first author manually annotated the corpora. All problematic instances were discussed with the other authors until agreement was reached. The sequence of annotations included two consecutive queries. The annotation process started with identification and tagging of all discourse entities. Subsequently, queries were analyzed in order to determine which entities exhibit an information change. This involved investigating both surface features and discourse phenomena which contribute to the identification of the topic development type.

4.1. Corpus Collection

The first dataset we used to create the corpus is a collection of pairs of queries that are spontaneously input by the query issuers. These sessions were mined by extracting sequences of queries (within a short time lapse) that matched a set of regular expression patterns (e.g. which capture patterns that started with 'wh'-word or 'how' phrases) or included pronominal mentions. Extracted queries only contained raw text and only automatically anonymized query pairs were available to the annotators.

Natural occurring query sessions are generally poor in discourse phenomena since human interaction to machines are linguistically under represented. This is because query issuers tend to formulate their questions in a way that minimizes usage of complex linguistic phenomena, while maximizing redundant language which may be unnatural in human-to-human conversation. In particular, pronominalization and anaphoric relations as well as sluicing and other types of ellipsis are less evident than expected in natural human-to-human conversation. While we do not observe a significant presence of these phenomena at present, we expect the language observed in queries to adapt to the machine becoming more reliable in understanding and showing the ability to generate such phenomena. In order to study these phenomena we also included query sessions from a semi-synthetic dataset to our pilot corpus.

This second dataset was collected by extracting sequences of queries without any constraint on the interrelatedness of their semantic and syntactic content. The extraction process for this set was the same as the one that is described for the first dataset. The data was then given to a second set annotators (native English speakers who are trained linguists) to use as the starting point to create sessions with certain

characteristics. The annotators revised extracted sequences of related queries or used the initial queries as inspiration and simply added a possible follow up query they invented. They were instructed to include in the query sessions specific phenomena that would naturally occur in conversation, such as anaphoric references and coreferential mentions.
The final pilot corpus we used in this study consists of 200 query sessions with the following distribution:

- 100 randomly sampled query sessions from the first naturally occurring dataset
- 100 randomly sampled query sessions from the second semi-synthetic dataset

4.2. Annotation Analysis

We have made several observations throughout our corpus analysis, which we will be presenting in this section. Specifically, personal pronouns were found to play a vital role in TEL, because they represent a natural way of keeping and exploring the same discourse entity (i.e. Participant) and, when annotated with this topic development tag they represent the constant element of the ISQ:

(4.1) **Q_1:** Who is$_F$ ***professor McGonagall***$_T$?
Q_2: How old is$_F$ ***she***$_T$?

In addition, the topicalized (constant) entity is commonly Predicate, for the cases of retaining the same topical entity and extending it to another participant, as in TEP:

(4.2) **Q_1:** How many goals ***did*** Beckham$_F$ ***score***$_T$ last year?
Q_2: How many goals ***did*** Zlatan$_F$ ***score***$_T$ last year?

We believe that patterns of interdependence between topic progression types and discourse entities can be identified in line with our analysis. This aspect could be further explored in future studies.
Moreover, a number of challenges have emerged in the annotation process. First, it should be noted that we carried out an additional search in the two datasets of consecutive ISQs specifically to find questions that contain auxiliary verbs, in order to investigate whether they are used for a specific conversational purpose, such as prefacing a question:

(4.3) **Q_1:** ***Prada shoes***$_T$
Q_2: Can I buy$_F$ ***them***$_T$ in Milan?

However, only 33 modal auxiliaries were found in the our datasets with just four in the sentence-initial position.[4] Other instances were all preceded by a 'wh'-word, as in (4.4)-Q_2 below:

(4.4) **Q_1:** What ***makes***$_T$ coffee mate$_F$?
A: Nestle.
Q_2: Where ***can I buy***$_T$ it$_F$ online?

Due to the non-entailed character of modal auxiliaries, questions that include these elements do not primarily ask about when and how an activity took place, but whether it took/can take place at all. However, from the point of view of the issuer's intention, which is of our primary interest here, the Topic of Q_1 shifts from the identity of the manufacturer to the activity of buying in Q_2. An agentive question such as: "Where do they sell it online?" would have the same purpose (and, in all likelihood, achieve the same goal). In fact, even in the queries where the modal/auxiliary verb contributes to the more conversational nature of the question (e.g. example (4.3)), the issuer's goal can be hypothesized to be basically the same. We decided to treat predicates including modal verbs (and all auxiliaries in general) as the other Predicate types (consequently, (4.4) is an instance of TSA from making coffee to buying it).
Another group of queries that were challenging to analyze are those where the topic development can only be specified via the answer:[5]

(4.5) **Q_1:** Who created the song The Edge of Glory?
A: Lady Gaga
Q_2: How old is$_F$ ***she***$_T$?

We regard (4.5)-Q_1 as an instance of an implicit Topic query, because it asks a question about the identity of an unfamiliar participant. In isolation, the title of the song would be a likely candidate for the topicalized entity; however, coreferential character of the constant element, Lady Gaga (implicitly present in (4.5)-Q_1, overtly expressed in (4.5)-A, and anaphorically picked up in (4.5)-Q_2) can be established via 'what-about'-anaphora. Since a query pair is, in the majority of cases, sufficient to identify the issuer's intent, we resorted to analyzing the context of the queries. We suggest this limitation of our study can be addressed in future research; however, we believe that our approach may in fact be useful for and facilitate answer retrieval thanks to identifying those discourse entities that are prominent to the issuer's goals.
A potentially problematic area for the ISQ interpretation are discourse/semantic phenomena and their significance topic development, for instance:

(4.6) **Q_1:** Does the Earth$_F$ ***rotate***$_T$?
Q_2: Does the Moon$_F$ ***rotate***$_T$?

This query was annotated as an instance of TEP (cf. (3.4) above); however, given the co-meronymic relationship of the Participants, this pair could be regarded as an instance of TEL. Likewise, future work should investigate whether positing such relationships results in devising a more accurate methodology in disambiguating topic development type.
Finally, another subset of queries where the interpretation was open to errors were the ISQs requiring world knowledge for disambiguation. This made the analysis of the topic development quite difficult, for instance (the query asking about the title of the American television series):

(4.7) **Q:** How to get away with murder?

[4] Remaining instances were found either in unrelated queries or in repetitive queries.

[5] Answers to each ISQ were available in the majority of the semi-synthetic queries.

Development Type	Frequency Count
Topic Exploration (TEL)	**80**
Topic Extension	**76**
Topic Extension Circumstances (TEC)	11
Topic Extension Participant (TEP)	46
Topic Extension Activity (TEA)	19
Topic Shift	**44**
Topic Shift Activity (TSA)	14
Topic Shift Participant (TSP)	30

Table 2: Number of occurrences of each topic development type in annotated datasets.

Discourse Entity Type	Q_1	Q_2
Participants	**104**	**83**
Agentive Participant (APTP)	4	1
Non-agentive Participant (NAPTP)	100	82
Predicates	**77**	**88**
Action Predicate (APRED)	25	37
Non-action Predicate (NPRED)	27	35
Stative Predicate (SPRED)	19	8
Procedural (PPRED)	6	8
Properties (PROP)	**19**	**29**

Table 3: Number of occurrences of each discourse entity type in the first and follow up queries.

4.3. Annotation Statistics

In addition to our above mentioned observations, we also obtained some statistics from the annotated corpora (see Table 2 and 3). The number of annotated ISQ sessions was low, however we tried to identify meaningful patterns. Overall, Table 2 illustrates that the majority of analyzed ISQs (nearly 80% in both datasets) is intended to either explore the same Topic, which is the case in Topic Exploration or retain the Topic and extend it with a new entity or circumstance (the case of Topic Extension).

We believe that this result illustrates the issuer's strategy in a query session of consecutive questions, as it shows that the issuer is more commonly interested in finding out more about the same topic than switching to another topic. Our category of Topic Shift comprise queries which involve related discourse entities (see Table 1 and section 3.3.3 above). This result also suggest that our topic progression annotation methodology is very practical from the point of view of capturing the conversational strategy that the issuer adopts, which is a preference to stay on the same topic within short ISQ sessions. A study including longer sessions can compare how/whether this tendency changes.

5. Conclusion

In this study, we present a concise and concrete annotation framework to tag discourse entities and topic development on a corpus of information seeking query pairs. An ISQ corpus annotated in line with this framework can be used as training data for dialogue management and computational semantics systems. One of the main ideas explored here is the relationship between the roles of discourse entities in ISQ sessions and topic development types. As discussed, the entities that constitute relationally given and constant information, can be regarded as the Topic of the queries. This is independent of referential definiteness/specificity, which the examples of Topic extensions illustrated. We believe this idea to be a particularly fruitful approach to model the conversational strategies that query issuers adopt while interacting with dialogue management systems, since it potentially delivers a fine distribution of the annotated phenomena.

6. References

Ariel, M. (2001). Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects*, 8:29–87.

Bos, J. (2008). Wide-coverage semantic analysis with boxer. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, pages 277–286. Association for Computational Linguistics.

Chafe, W. (1994). Discourse, consciousness, and time. *Discourse*, 2(1).

Chai, J. Y. and Jin, R. (2004). Discourse structure for context question answering. In *Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL*, volume 2004, pages 23–30.

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Gundel, J. K. and Fretheim, T. (2004). Topic and focus. *The handbook of pragmatics*, 175:196.

Heim, I. (1982). The semantics of definite and indefinite noun phrases.

Huddleston, R., Pullum, G. K., et al. (2002). The cambridge grammar of english. *Language. Cambridge: Cambridge University Press*, pages 1–23.

Jackendoff, R. and Culicover, P. W. (2003). The semantic basis of control in english. *Language*, 79(3):517–556.

Kratzer, A. (2000). Building statives. In *Annual Meeting of the Berkeley Linguistics Society*, volume 26, pages 385–399.

Lambrecht, K. (1996). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*, volume 71. Cambridge university press.

Prince, E. F. (1992). The zpg letter: Subjects, definiteness, and information-status. *Discourse description: diverse analyses of a fund raising text*, pages 295–325.

Strawson, P. F. (1964). Identifying reference and truth-values. *Theoria*, 30(2):96–118.