

Information extraction for multi-participant, task-oriented, synchronous, computer-mediated communication: a corpus study of chat data*

Cassandre Creswell, Nicholas Schwartzmyer, and Rohini Srihari

Janya, Inc.

Amherst NY USA 14228

{ccreswell, nschwartzmyer, rohini}@janyainc.com

Abstract

Applications for synchronous computer-mediated communication, i.e. chat, like instant messaging and chatroom channels, are playing an ever-increasing role in task-oriented (vs. recreational) domains. Information extraction (IE) from chat could provide great value to any process where acting on a real-time information stream is important. However, no previous work exists on performing IE on chat data, and only limited research has been done on the linguistic differences between chat, spoken dialog, and written text. Chat likely poses several challenges for standard IE methods developed for heavily-edited written text, including: (i) surface-form noise, e.g. non-standard usage of punctuation; (ii) discourse-level noise, e.g. complex discourse structures that make resolution of the high frequency of context-dependent and anaphoric linguistic forms even more difficult. This paper describes an annotated corpus of task-oriented chat logs created in order to assess how the noise in chat data will affect the development of high accuracy IE technology for chat.

1 Introduction

The goal of information extraction (IE) is to automatically identify entities, events, and relationships of interest in natural language content. IE as it exists today evolved from the work and ideas of the Message Understanding Conferences (MUC) which began in the late 1980s. The MUC tasks, although tending to increase in complexity every year, consistently used single-author written reports or read news as data sets [Grishman and Sundheim, 1996]. As a result, most IE techniques have been developed with relatively static, one-sided discourse genres as input.

The Automatic Content Extraction (ACE) program, the successor of MUC, has sought to rectify this somewhat by steadily adding training and evaluation data of a variety of discourse types, including transcripts of broadcast and telephone conversations, usenet archives, and weblogs [ACE,

2005]. The inclusion of such data is an acknowledgment of the need to address the problems that other types of natural language discourse pose for IE. However, the nature of these problems remains for the most part unexamined.

The ever-increasing popularity of synchronous computer-mediated communication (CMC) channels like instant messaging (IM) and chatrooms in the workplace, including both industry and military settings, is potentially a rich new input source for IE applications. Given real-time extraction software, so-called CHAT data could, for example, automatically provide up-to-the-minute information about events and scenarios to decision makers, continuously enriching their knowledge bases and allowing them to act more quickly.

A key property of chat that distinguishes it from the standard written text that IE software has been designed for is its noisiness. The sources of noise in chat in addition to surface noise are many. First, it is a dynamic form of discourse; as a chat discourse progresses, the propositions expressed in it may be contradicted and revised. Second, it is interactive. Chat necessarily has contributions from more than one participant, allowing for disagreement, misunderstanding, and, simultaneously, a large degree of implicit shared knowledge. These properties make chat very similar to spoken dialog, another genre where relatively little IE work has been applied. Third, in contrast to most of spoken dialog, the ordering of turns in chat is less constrained. Because past utterances are still accessible visually, it is easier to have multiple threads of discussion being conducted simultaneously, and turns from unrelated threads may be interleaved. This makes automatically detecting underlying discourse structure(s) of chat data a very difficult task. As a result, resolving the semantics of forms that are dependent on local discourse context is also more challenging.

In the remainder of this paper we will first describe a corpus of task-oriented chat data annotated for noise and noise-sensitive properties. We will then describe the annotation schema applied to the corpus in order to study the quantitative and qualitative differences between chat data, written discourse and spoken dialog. Finally, we present examples of the various sources of noise in chat data and discuss the problems they potentially present for automatically extracting information about entities and events from chat data.

* This work was supported in part by SBIR grant FA8750-06-C-0157 from the Air Force Research Laboratory (AFRL)/IFED.

2 Task-oriented chat data: the GeoTools corpus

In contrast to the majority of previous corpus-based work on synchronous CMC¹, in our corpus the participants are engaged in communication with the purpose of exchanging information on a highly-complex collaborative task, rather than engaged in communication for primarily recreational or social ends. Task-oriented dialogs have played a significant role in psycholinguistics and computational linguistics in shedding light on discourse-level coordination of meaning between speakers [Brennan, 2000]. We believe that a focus on task-oriented interactions can have a similar impact in the study of chat data. In addition, in terms of technological impact, IE is likely to be more important for task-oriented than recreational domains.

The GeoTools chat corpus is comprised of 56 IRC logs for the GeoTools project, an open source Java GIS toolkit [GEOTOOLS, 2005]; the logs were downloaded from <http://geotools.codehaus.org/IRC+Logs>). Altogether the corpus contains ~210K words with about 15,000 participant turns and spans from about April 2004 to December 2005. The purpose of the interactions is a weekly meeting of GeoTools software developers, around five per session, in which they discuss issues in the development process relevant at that current time, typically the delegation of work, the status of subtasks, directions in which the projects can be taken, or problems that a particular developer is having with a previously assigned task. On a whole, the interactions have a structure similar to an informal business meeting, where an agenda is given at the onset, and each subtopic is discussed with some degree of detail. Typically, the participant who posts the agenda manages the transition between subtopics on the agenda.

As a corpus of task-oriented interactions, the GeoTools corpus is an appropriate model for other on-the-job chat domains. By TASK-ORIENTED, we mean specifically that the structure and purpose of the discourse is guided primarily by extra-linguistic objectives. That is, the participants are engaged in a shared task which is entirely independent of their communicative interaction, and the purpose of communicating is to achieve goals with respect to that task. The GeoTools logs have this exact property, making it a better model for chat data in workplace applications, than interactions on a typical recreational IRC application. For example, in task-oriented chat, one to a few participants take command of the floor, contributing the most messages and thus shaping the conversation at hand, typically toward goal realization [Birnholtz *et al.*, 2005]. In a recreational domain, however, there is less need for adherence to a small set of topics or discourse purposes because participants will rarely seek to achieve a more specific shared goal than to be generally cooperative.

¹A notable exception being Ivanovic’s work on instant messaging (IM) conversations between customers and customer-service representatives [Ivanovic, 2005b,a]. These papers, however, only describe two-person conversations, rather than the multi-party conversations in our corpus.

3 Annotation schema

Our annotation schema consists of two parts. The first, described in Section 3.1 was developed in order to investigate the extent and type of surface-level noise likely to interfere with the core natural language processing modules in our information extraction system (e.g. part-of-speech tagging, lexical lookup, NE tagging, parsing). The second, in section 3.2 was intended to allow us to assess how the complexities of discourse-level properties in chat will affect the extraction of events and entities from such data.

3.1 Surface-form phenomena

Chat data potentially contains a high frequency of linguistic properties that pose a challenge to the performance of IE software developed primarily for professionally-edited, narrative text. In an informal, relatively-unedited medium like chat, produced under time constraints, errors and non-standard usage, whether in spelling, punctuation, orthography, or grammar, are common. Content extraction depends on accurate detection of patterns in linguistic data. Therefore, even slight variations in spelling, orthography, punctuation, and grammar can have a large impact on software designed for “clean” text that does not display such variation. In light of this, we annotated a subset of the GeoTools corpus (23 of 56 logs, 6788 turns) for four properties: non-standard orthography, non-standard punctuation, misspellings, and ungrammatical constructions. In addition, because we were interested in the discourse-level phenomena of dialog act types, we also marked up the corpus for turn-final punctuation.

Non-standard orthography includes both lowercasing in contexts where capitalization is expected and vice versa—whether it is sentence-initially, as in (1) or as the first character of a proper noun or in a word that normally takes mixed case (e.g. *geotools* for *GeoTools*) or all capitals (e.g. *I* for *I*).

- (1) Martin that’s not what I had in mind.

A turn-initial word was deemed not necessarily missing capitalization unless it began a sentence² because one sentence may span several turns. Almost half of all turns had an instance of non-standard orthography; about 60% of these were cases of upper case appearing as lower.

Non-standard punctuation, present in about 45% of turns, includes two phenomena: incorrect usage of punctuation and omission of punctuation. We did not wish to include all possible violations of standard punctuation usage, but only those that could have a significant effect on the ability to parse the grammatical structure of the text. For example, if a fragment were encountered where usage guidelines would prescribe a semicolon, but a comma is encountered instead, this was to be ignored. In contrast, if a sentential unit ended with punctuation which is not expected sentence finally, this was to be marked. Similarly, if a sentence-final punctuation was found in a position that it should not (e.g. a period in place of a comma), this was also marked.

With respect to the annotation of omitted punctuation, the presence of punctuation is often very important in designing

²We define a TURN as every instance of a speaker entering text and then pressing return.

systems for high-accuracy, high-recall IE. Missing sentence-final punctuation was always marked, but because the schema required the annotator to supply the correct punctuation, there was the possibility of some difficult annotation judgments. In such cases, we used a set of heuristics to decide if an omitted punctuation mark would have significance for interpreting clause structure. The most broadly applicable of these was that if a sentence could be interpreted as a declarative, then it should have a period as its final punctuation.

Turn-final punctuation may provide valuable information about dialog act information and, hence, discourse relations between utterances. The guidelines for annotating this category were quite simple: annotate any punctuation that appears turn finally. The exceptions were right parentheses, where any punctuation inside them were to be considered final, and ellipsis, where regardless of spacing all elements of the sequence were to be considered final.

Misspellings make recognition of the most fundamental linguistic units of text difficult because the majority of lexicons and grammars will not be able to match strings that differ by more than capitalization differences. A misspelled word will either be mistakenly recognized as a different word or as an out-of-vocabulary token. In annotating misspellings we assumed standard American English spellings. In determining whether an error was a misspelling vs. an ungrammatical construction, such as a violation of subject-verb agreement, the annotator was to choose ungrammatical construction, as in *He work at Sun Microsystems*. Misspellings occurred in about 13% of annotated turns.

Ungrammatical constructions were defined to be any multi-constituent sequence of words which did not conform to standard written American English usage conventions, as exemplified by newswire text expected as input to most commercial IE systems. This, then, encompassed two phenomena. The first is constructions that are not allowed by the rules of English syntax. This group would include things such as subject-verb agreement violations, constituent ordering violations, etc. The second group is what we shall call non-standard usage constructions (NUCs). NUCs do not violate the rules of English grammar but are not standard features of standard written English, possibly because of medium and/or register effects. Examples of the latter are forms that have undergone phonological reduction, and their spelling is intended to reflect this (eg. *kinda, gotcha, dontcha, dunno*, etc.). A set of heuristics were formulated to specify what text span should be marked as part of an ungrammatical construction. For example, any ungrammaticality involving the main verb was taken as having scope over the entire sentence. Ungrammatical constructions were somewhat more frequent than misspellings, appearing in almost 18% of turns.

3.2 Discourse-level phenomena

Discourse-level properties annotated in the GeoTools corpus included inferential phenomena related to several syntactic types: verb phrases, noun phrases (including time expressions and location deictics), and sentences/utterances. Each of these constituent types present complexities of interpretation that are either unique to chat data or of greater significance in this near-synchronous, informal, interactive mode

of communication than in other written text. Our annotation guidelines built on existing annotation guidelines, including the DRAMA manual for coreference of NPs [Passonneau, 1996], the DAMSL and MRDA standards for dialog act tags [Jurafsky *et al.*, 1997; Dhillon *et al.*, 2004], and the Penn Discourse Treebank for dependencies (semantic relations) between utterances [The PDTB Research Group, 2006].

Noun phrases

We annotated all NPs in 20% of the corpus. Like other discourse-level phenomena, NP anaphora resolution is a challenging problem for IE and will likely be more complicated in chat data. For instance, turns may be taken out of order, meaning that anaphoric noun phrases (NPs) may show some very complicated resolution patterns, necessitating thorough analysis of current coreference resolution strategies. Our annotation schema for these NPs follows the DRAMA annotation schema [Passonneau, 1996], an important precursor for other NP annotation schemas [Poesio, 2004]. Here we concentrate on describing the differences between the DRAMA schema and our own. Analysis of our NP annotations is still on-going; we are most interested in how chains of coreferent NPs interact with chains of dependent utterances; the latter is defined below. Also of interest will be NPs like pronouns and definite NPs that typically have explicit NP antecedents in written, formal text but may lack them in chat.

The supercategory NP was divided into the following: REFERENTIAL and NON-REFERENTIAL. We defined a referential NP as any NP which refers to (i.e. picks out) a discourse entity within the discourse model. An expression is annotated as a referential NP if it introduces a new discourse entity or if it refers back to a previous discourse entity [Passonneau, 1996, pg. 5]. Referential NPs can then be further subdivided into the following: REGULAR and DISCOURSE DEICTICS [Webber, 1991]. An NP of either of these types can be coreferential with another NP, and coreferring NPs that occur before an NP in the discourse are its antecedents. Discourse deictics, however, have as their textual antecedent VPs or sentences rather than NPs, as shown in (2), where the entire previous sentence is the antecedent for *this*.

- (2) **The tool will be available tomorrow. This** means we need to do bug fixes tonight.

All other referential NPs are regular NPs, including: (any non-discourse deictic uses of) personal pronouns, non-restrictive relative pronouns, proper names, free relatives, zero/empty pronouns³, interrogative pronouns, and common noun-headed NPs. We also annotated gerund participles, but these were only marked if they fulfilled the following conditions: (i) presence of a determiner (including possessive pronoun) or an adjectival modifier or (ii) a plural ending. This is a smaller subset of gerundives than included in the DRAMA schema. Another difference with DRAMA is the annotation of deictic adverbs, e.g. *here* and *today*. These were annotated

³For example, this includes the subject of imperatives (*Please go ahead and...*) and the subject-drop found in less formal registers of English (*Did it./Could've been me.*) In cases of two VPs conjoined by a coordinating conjunction, we assume the second VP shares the subject of the first and does not include a missing second subject.

by us separately as location deictics and time expressions. DRAMA also excludes possessive pronouns, appositive NPs, and negated noun phrases. We, in contrast, include all of these categories. A special feature of chat data is that the identity of the speaker appears at the beginning of the speaker's contribution to the discourse. As such, we included the speaker's username in the set of referential NPs; this makes it possible to link indexical personal pronouns like *you* and *I* into a chain of coreferring NPs.

The subcategory NON-REFERENTIAL includes any NPs not used to pick out a discourse referent, including NPs found in discourse connective phrases (*on the other hand*), pleonastic and weather *it* (*It seems that John has left/It's raining*), and idiomatic uses of nouns, such as *thumb in rule of thumb*. The heuristic for determining whether an NP is non-referential was to check whether it would be possible to use a pronoun or other NP to refer to it again. If it appeared impossible, then it was likely a non-referential use (e.g. *on the other hand...*/??that hand*).

Every NP was marked as one of three types NON-REFERENTIAL, NON-NP ANTECEDENT (=discourse deictics), and ANAPHORIC(=regular). If an NP had any preceding NP in the discourse with which it was coreferential, it was linked back to it with a COREFER link. Because all coreferential links can be collapsed using transitive closure to create a set of coreferring NPs, it does not matter whether an NP is linked to its closest coreferring NP. It simply needs to be linked to at least one preceding, coreferring NP.

Verb phrase ellipsis

Verb phrase ellipsis (VPE) is the omission of the lexical verb head, its arguments, and its modifiers in a verb phrase. In the absence of the infinitive marker or auxiliary, the pro-form *do* is used. The semantics of the omitted syntactic material must be recovered from context. Two examples of VPE appear in (3–4), where the boldface *to* and *did* are all the form left to indicate the content, *leave soon* and *stay late*, respectively.

- (3) Although Max thinks I'll leave soon, I don't want **to**.
 (4) Although Max didn't intend to stay late, he **did** anyway.

VP ellipsis is again a challenging problem in IE because its semantic content depends on discourse-level inference and interpretation. This is problematic in chat data because determining what should count as the appropriate discourse context for resolving the ellipsis is complicated by the possibility of multiple threads and lack of strictly ordered turns. As such, previous work on the topic (Hardt [1997]; Nielsen [2003]) may not be directly applicable. However, it appears that VPE was relatively rare in the GeoTools corpus, with only 61 instances across all logs. In addition, only 53 of these 61 had an antecedent beyond a turn boundary. This could greatly simplify VPE resolution in chat data.

Sentential and non-sentential utterances

Our utterance-level schema is comprised of three annotation activities, identification of utterance units, assigning dialog act tags to utterance units, and linking each utterance unit to the one it depends on. It was applied to about 20% of the corpus, or ~ 2700 turns, which encompassed ~ 3900 utterances.

Utterance units can be either sentential or non-sentential. Identifying utterance units has much in common with the marking of sentence boundaries in automatic speech recognition output. This task is important for several reasons. Because of the dynamic and informal nature of chat data, sentence boundaries are less likely to be marked with the appropriate (or any) punctuation, as shown in (5), than in static written genres where the author or some third party edits the text before it is presented to its intended audience. As such, sentence-final punctuation cannot be used by itself to identify utterance units.

- (5) jgarnett I would love a repalcement [sic], I know we went to a lot of work though...

In addition, not every discourse contribution (or TURN) will correspond to the typical syntactic definition of a sentence or functionally-independent clause. For example, an isolated noun phrase may serve as a complete utterance when a participant answers a question. Utterance boundaries may also not correspond to turn boundaries. Turn boundaries are points where a participant inserts a carriage return, making their contribution readable to the other chat participants. Multiple utterances may appear in a single turn, and, a single utterance may potentially span multiple turns.

Sentence/utterance boundary information can be used in information extraction for parsing. Additionally, it is important for coreference resolution in order to determine the saliency of potential NP antecedents. It is also crucial information for resolving NPs with non-NP antecedents, mentioned above as DISCOURSE DEICTICS. In addition, in order to assign even a very simplified discourse structure to a chat transcript, i.e. link related content that crosses turn boundaries, the basic building blocks of the structure, that is utterance units, must be assigned.

Once utterance units were marked, each utterance was assigned a dialog act tag, e.g. AFFIRMATIVE ANSWER, YES-NO QUESTION, THANKING, ACTION-DIRECTIVE, etc. In written newswire or other similar genres, assigning dialog acts is of limited utility for information extraction because dialog acts other than stating information are infrequent. In contrast, in multi-party genres like conversational speech or computer-mediated chat, dialog acts have a much greater informational role. The factuality of a contribution will depend on whether it appears in a statement vs. in a question or command. Linking related turns to each other can be improved by identification of dialog act tags because certain sequences of tags are predictable. For example, with a high probability, questions are followed by answers, and action-directives are followed by acceptances or rejections. The dialog act tag set draws heavily on the annotation manuals from previous efforts in the literature [Jurafsky *et al.*, 1997; Dhillon *et al.*, 2004].

The final component of utterance-level annotation is linking each utterance to the previous utterance it DEPENDS on. We defined the DEPEND relation between two utterances U_1 and U_2 as one of the following three conditions holding: (i) U_2 is the paired dialog act for the dialog act for U_1 or (ii) U_2 stands in one of the seven implicit discourse relations used by The PDTB Research Group [2006] with U_1 or (iii) U_2 is

marked with a discourse connective whose first argument is U_1 .

We did not define all possible dialog act pairs, but instead illustrated the concept by providing examples: Statement + Accept; Yes-no Question + Negative Answer; Statement + Correction; Offer-Suggest + Commit; Commit + Thanking; Thanking + Downplayer.

In some cases, an utterance will not depend on any other utterance, and so it will not be linked in a dependency relation. This category includes discourse-initial utterances, paranthetical utterances, uninterpretable/abandoned utterances, or utterances that begin an entirely new topic.

An important subtype of utterance in chat data, as in spoken dialog, are NON-SENTENTIAL UTTERANCES (NSUs) [Ginzburg and Fernandez, 2002]. Their non-sentential property was annotated as a boolean feature, but otherwise they were treated in the same way as sentential utterances with respect to being assigned dialog act tags and dependency links. This required us to unify the dialog act tag sets of Jurafsky *et al.* [1997] and Dhillon *et al.* [2004] with Ginzburg and Fernandez [2002]’s taxonomy of NSUs. The importance of NSUs in chat data are that they are not semantically complete propositions. In order to resolve their missing content—a prerequisite for extracting such content—it will be useful to locate their antecedent. In our annotation, this is the equivalent to the utterance they depend on. We expect that identifying and resolving NSUs may be an important task in IE for chat. In our corpus, of the 701 NSUs, over half found their antecedent in the immediately preceding turn; this is encouraging evidence for the tractability of resolving NSUs in chat.

Figure 1 shows a sample utterance-level annotation of a chat transcript from the GeoTools corpus. Several phenomena are illustrated here. With respect to utterance boundaries, turn T7 contains two utterances (7,8). Utterance 11 spans two turns T10 and T11. Some utterances depend on each other by virtue of their dialog act pairing, for example, utterances 6 and 10 (wh-question and answer). Others are paired because of their content (12 and 11, which are in an implicit ADDITIONAL-INFO relation.) Some utterances do not depend on any others because, for example, they are initial openings (1,7) or they begin a new topic (4). An interesting case is utterance 3 which does not depend on any prior utterance. The speaker is using this utterance to check whether his next utterance will be felicitous, i.e. if the meeting has not commenced, then he should not introduce the meeting’s topic.

4 Sources of noise in chat data

From the perspective of information extraction software designed for edited, single-author written discourse, chat data contains considerable noise. By *noise* we mean any aspect of the data, including its surface form, that potentially interferes with extracting its true semantic content. This naturally includes the surface form phenomena discussed in Section 3.1, but we feel it applies equally well to the discourse-level phenomena to be discussed in this section. In the case of information extraction, the true semantic content of a text should correspond to the set of entities and events and their attributes

T1	jgarnett	[Hi guys] ₁ OP
T2	rschulz	[hi everyone] ₂ OP
T3	jgarnett	[I assume this is the meeting then?] ₃ DY
T4	jgarnett	[James is very close to the 2.0 release] ₄ SD
T5	jgarnett	[I told him we would go over the readme for him.] ₅ OS
T6	rschulz	[where is the readme] ₆ QW
T7	cholmes	[Hi everyone.] ₇ OP [Happy New Years!] ₈ OP
T8	jgarnett	[Hi chris!] ₉ OP
T9	jgarnett	[http://svn.geotools.org/geotools/branches/2.0.x/gt/README.txt] ₁₀ AA
T10	jgarnett	[My first thoughts are to replace the vague “provide support for specific data formats or sources, or provide specific additional functionality”
T11	jgarnett	with a breakdown of the actual modules and what they do.] ₁₁ SV
T12	jgarnett	[I kinda wish the readme was html :] ₁₂ SV

Dependencies: 1 → ∅; 2 → 1; 3 → ∅; 4 → ∅; 5 → 4; 6 → 5; 7 → ∅; 8 → 7; 9 → 8; 10 → 6; 11 → 5; 12 → 11.

Figure 1: Sample annotation of chat transcript with utterance boundaries (marked with square brackets) and dialog acts (following each utterance.) Dependencies are listed separately. Dialog act tag abbreviations used here are the following: OP, DY, SD=statement-non-opinion, OS=offer-suggest, QW=wh-question, AA=affirmative-answer, SV=statement-opinion.

described in the text⁴ In particular, there are three properties of chat that conflict with extracting this information (i) it is dynamic, (ii) it is interactive, and (iii) it allows for non-local context. In this section, we will illustrate all these properties relative to examples from the GeoTools corpus.

4.1 Dynamic content

Because chat data is dynamic, the information and beliefs of participants evolve and can be revised as the discourse progresses. Compared to most written genres, speech acts besides assertions (statements) such as questions, commands, requests, suggestions, corrections, and clarifications are more common in chat. This has several important consequences for IE because fewer mentioned entities, events, and attributes can be interpreted as factual. Examples (6–7) contain two illustrations of this issue from the GeoTools corpus. In (6), from the initial use of the NP *a set of geoapi interfaces we should be using instead* it is not clear whether the NP has a referent because it is inside a question about its existence. If the response to the question was *no, there’s no such set*, then a corresponding entity should not be created for the NP in the extracted content from this exchange. In this case, such a set does exist, as the subsequent turns make clear, referring to it with *geoapi interfaces* and *org.opengis.referencing*. In either case, the status of the NP’s referent can change as additional

⁴In some cases, then, what we are labeling as noise in chat data would not be noise if IE software were designed around chat rather than written newswire.

contributions are made to the discourse. This dynamic effect appears at the level of propositional content as well, as shown in (7). Here, in the first turn, jgarnett declares that it is unclear whether some proposition *P* hold or not. In his next turn, he changes his mind, and says that *P* is the case. The next speaker, polio, partially agrees, stating that *P* is probably the case.

(6)

jmacgill	should we be using org.geotools.referencing directly or is there a set of geoapi interfaces we should be using instead?
Martin	We should use geoapi interfaces
Martin	org.opengis.referencing

(7)

jgarnett	It is unclear from http://geoapi.sourceforge.net/snapshot/javadoc/org/opengis/referencing/cs/AxisDirection.html
jgarnett	if the referencing/cs package is intended to live or not
jgarnett	oh wait it is supposed to live it is org/opengis/cs that is trying to die.
polio	it probably is, since it's part of martin's new stuff

4.2 Interactive content

To the extent that chat is like spoken dialog, participants can take more information for granted in chat discourse because they have the opportunity to alert each other of misunderstandings. For example, if reference to a particular entity is unclear, a participant can ask *Which one?*—something that is not possible in a non-interactive written mode. Coordination of participants' beliefs, usually called GROUNDING [Clark, 1996], plays a large role in chat, like it does in spoken dialog. This is of particular importance in task-oriented chat where the need for participants to have an accurate set of shared beliefs in order to accomplish their task is crucial. Because grounding, clarification, and correction of misunderstandings are possible in an interactive mode of communication, participants can rely on each other to do more inference, allowing the use of content-poor forms like discourse deictics and VPE. This interactivity results in effects similar to those caused by the dynamic property of chat; the factuality of propositions and the reference of NPs can change, and in fact these two things are interdependent. The exchange in (8) illustrates this well. Speaker jgarnett does not understand the reference of rschulz's NP *the CRSServices proj issue*, asking *which one*. Based on his initial understanding, he provides a hedged answer to the question *I think we did*, where the VPE *did* means *get the CRSServices proj issue worked out*. After rschulz explains which issue he intends to refer to, jgarnett changes his answer to *ah no I never figured that out*, where the discourse deictic *that* refers to rschulz's explanation. So, in this exchange, the factuality of the event described by *we got the CRSServices proj issue worked out* changes depending on the reference of the definite NP, which changes as the participants continue to add information to the discourse. An IE system must take into account that the meaning of the text can change as the participants interact.

(8)

rschulz	also, did you get the CRSServices proj issue worked out?
rschulz	(Jody)
jgarnett	which one? I think we did - our wkt was not complete on the shapefile.prg
jgarnett	bleck - I mean shapefile.prj
rschulz	you had a test using crsservice to re-project geometries. It gave wierd results for the bounding box
rschulz	I thought crsservice might be modifying its input geometries
jgarnett	ah no I never figured that out

4.3 Multiple, interleaved topics

Because there is a visual, static record of their communication, participants in chat are not as constrained in their turn-taking as are participants in spoken conversation. As a result, discussing multiple topics, or threads, simultaneously is possible. In addition, in a mode (like IRC) where composition of a contribution is not visible, other participants are not entirely aware of when a new turn will be added, and so some accidental interleaving of unrelated turns takes place. This means that what counts as "local" or "salient" content for the purposes of resolving anaphora or ellipses, may not be adjacent. This is possible of course in spoken or written modes too. However, here the non-adjacency can be caused not only by embedded, but strictly not crossing, subsegments of discourse structure, but also by crossing dependencies. The interleaving of multiple topics is illustrated in Figure 2.⁵ Here, from turns 1 through 11, the participants are discussing agenda item 6, *postgis fid details*. Then, at turn 12, jgarnett mentions a side topic, i.e. that he is ready to make a new commit; this is not responded to until turn 17, and again in turns 22 and 23 where the commit is noted both by the CVS system bot (CIA-11) and jgarnett himself. Mixed with these turns, discussion related to the agenda item continues. Here in turn 14 polio responds to 11; his use of *that* refers to making *the fixes* referred to in 11. This leads to the following chain: [14,15,16,18,19,21,24], where, for example, the pronoun *he* in turn 19 resolves to the speaker jesse (the *i* in 15) and the pronoun *that* in 21 resolves to the entire content of turn 19. Another crossing dependency is turn 13, which is a response to 10; in 13 the missing grammatical object of *review* is coreferential with the instances of *it* in 10.

Clearly, any anaphora resolution algorithms which depend merely on simple proximity in the text in order to determine the salience of potential antecedents will be foiled by this type of interleaving. In addition, even an algorithm which allows for embedded discourse structures where referents can come in and out of focus will be challenged by this chat data. This is because a single discourse structure cannot be assigned; the different threads must be separated, and each thread assigned its own structure.

5 Future work: A chat-based IE engine

In this paper, we have described a corpus of chat data and annotation schema designed to mark up phenomena that are

⁵For simplicity's sake here, we will talk only about turns rather than utterances.

1 jgarnett ...6) fill me in on postgis fid details - I realized I have patches waiting, on somewhat related stuff, that I want to integrate today me = cholmes Well jgarnett != cholmes but me in the subject = cholmes

2 jeichar I'm just tracking down the change I made to make it work.

3 jgarnett Last week was so long ago.

4 cholmes Yeah, just curious what you guys did, as I realized I have patches to apply, that may be similar. Sorry I forgot about them when you guys were working on it.

5 jgarnett In that case we may have just duplicated work - Postgis was not generating its FIDs in any sane manner. Their was a class their that did FID generation but it was not being used. And the create loop did not bother to return the FID even when it was known. A simple mistake really .. but one that cost brain cells to debug.

6 jeichar That's more or less it.

7 CIA-11 dzwiers * r11395 geotools/gt/ (4 files in 4 dirs): JTS update ... new env

8 jeichar A FID was created but not set in the feature so it was lost immediately after creation.

9 cholmes Do you remember what class the fixes are in?

10 jeichar I'm trying to find it. I'd love for you to check it.

11 cholmes (actually it looks like the fixes are related to createSchema, so it may be a different problem).

12 jgarnett While we wait dblasby I almost have the "empty" hsql datastore ready to commit.

13 cholmes Cool, I'd be happy to review.

14 polio ah, that was me

15 jesse and I added crs support to postgis

16 cholmes Successfully?

17 dblasby thanks jody

18 cholmes Where did you do the changes?

19 polio he made a crs factory thingy and I hooked it into postgis

20 jmacgill sorry, gotta run

21 polio that was the createschema changes ... it was just a couple lines that called the PostGIS AuthorityFactory that allowed the featureType to know it's projection

22 CIA-11 jgarnett * r11396 geotools/gt/ext/hsql/tests/: Removed file/folder

23 jgarnett * r11397 geotools/gt/ext/hsql/ (5 files in 5 dirs): Removed file/folder

24 cholmes Ok, cool.

Figure 2: Sample chat log excerpt with multiple interleaved threads.

likely to present difficulties for applying information extraction software successfully to chat data. The corpus annotation has allowed us to examine patterns of surface sources of noise, such as misspellings and non-standard usages of orthography, punctuation, and grammar. Examining discourse-level properties was intended to elucidate more complex sources of noise like the interleaving of multiple topics and the effects of a dynamic, interactive mode of discourse where semantic content changes as the discourse progresses.

As future work, we are developing an IE system with an architecture specifically designed to process chat data, for which the corpus study has provided us with design constraints posed by the complexity of chat. We provide a very brief description of the design here. Our implementation of a chat-based IE system will be an augmented version of our existing IE software, Semantex⁶. The key changes to the linguistic processing components of the software will be made to deal with the two sources of noise described in this paper. In order for the system to perform robustly when faced with the surface noise typical of chat data described above—non-standard orthography (case), punctuation, spelling, and grammar—we will build on solutions already incorporated into the system for other noisy data. This means pursuing a combined approach of restoring standard usage when possible and using noise-resistant processing elsewhere. An example of the former is the existing ability to restore standard case to all uppercase or lowercase documents [Niu *et al.*, 2004]. An example of the latter is Semantex's parsing capabilities which do not expect to produce a fully-connected parse tree for every sentence, and, as such, can easily cope with sentence fragments.

To cope with the complexity of discourse structure, rather than having a single pipeline of linguistic processing levels that apply to an entire document in succession, processing will be split into sentence-level and discourse-level modules. The former will apply once to every turn added to the chat log and will consist of any processing that does *not* require contextual information beyond sentence-boundaries.

Document-level processing will allow for lookback into any content processed earlier in order to make processing decisions about the current chunk of text, for example in anaphora resolution. The discourse-level processing will have two modes. Initially in the development process, it will be run repeatedly on a growing set of turns. Later, however, with the addition to the system of a more abstract representation of document content, only the key semantic content from each turn will need to be retained to be used in the document-level processing of new content. This will allow a more efficient incremental process at the discourse-level.

With either of these modes, we will allow the transcript contents to be reordered or linked in a way that represents multiple threads present in a single chat transcript. This is intended to provide a source of non-local content to draw from when resolving anaphoric phenomena, that is con-

⁶Semantex is a domain-independent and domain-portable IE engine, derived from an earlier IE engine known as InfoXtract . A detailed description of the system is available here: Srihari *et al.* [2006]

tent not immediately adjacent in the transcript. We will explore a combination of two methods of organizing chat utterances into topical-intentional threads, (i) unsupervised lexical-based clustering of turns and (ii) supervised learning for linking turns. Each method has precedents in the literature. Zhou and Hovy [2005] use unsupervised clustering to identify threads for summarizing IRC chats. They also use supervised learning methods to identify adjacency pairs in their chat logs. The latter technique is similar to that used by Galley *et al.* [2004] for identifying agreement and disagreement responses and by Schlangen [2005] for identifying antecedents of NSU, in both cases for multi-party dialog.

Overall, the problem of high-level information extraction from chat data (for example, the compilation of multiple entity mentions into a single entity profile or the compilation of event attributes from multiple mentions) is extremely hard, owing to the noisy nature of chat data. Therefore, we expect initially that the most impressive results of applying IE to chat may be the rapid generation of alerts when objects of interest appear at the mention level in the output of shallow parsing. Ultimately however, leveraging the value of chat in time-sensitive applications will require extracting better and more complete information. This in turn will require discourse-level processing, as important facts about entities of interest may only be mentioned using pronouns or other syntactic constructions (definite NPs, VPEs, and NSUs) that require decoding at the discourse level.

References

- ACE. Project specifications: ACE data overview, 2005.
- Jeremy P. Birnholtz, Thomas A. Finholt, Daniel B. Horn, and Sung Joo Bae. Grounding needs: achieving common ground via lightweight chat in large, distributed, ad-hoc groups. In *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 21–30, New York, NY, USA, 2005. ACM Press.
- Susan Brennan. Processes that shape conversation and their implications for computational linguistics. In *Proceedings of the 38th Annual Meeting of the ACL*, Hong Kong, 2000.
- H. H. Clark. *Using Language*. Cambridge University Press, Cambridge, UK, 1996.
- Rajdip Dhillon, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. Meeting recorder project: Dialog act labeling guide. Technical Report TR-04-002, ICSI, February 2004.
- Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 669–676, Barcelona, Spain, July 2004.
- GEOTOOLS. Geotools, the open source Java GIS toolkit, December 2005.
- Jonathan Ginzburg and Raquel Fernandez. A corpus study of non-sentential utterances in dialogue. *Traitement Automatique de Langues*, 43(2):12–43, 2002.
- Ralph Grishman and Beth Sundheim. Message understanding conference-6: A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, 1996.
- Daniel Hardt. An empirical approach to VP ellipsis. *Computational Linguistics*, 23(4):525–541, 1997.
- Edward Ivanovic. Automatic utterance segmentation in instant messaging dialogue. In *Proceedings of the Australasian Language Technology Workshop*, pages 241–249, Sydney, NSW, Australia, December 2005. Australasian Language Technology Association.
- Edward Ivanovic. Dialogue act tagging for instant messaging chat sessions. In *Proceedings of the ACL Student Research Workshop*, pages 79–84, Ann Arbor, MI, June 2005.
- D. Jurafsky, E. Shriberg, and D. Biasca. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Technical Report 97-02, Institute of Cognitive Science, University of Colorado, Boulder, 1997. Draft 13.
- Leif Arda Nielsen. A corpus-based study of Verb Phrase Ellipsis. In *Proceedings of the 6th Annual CLUK Research Colloquium*, pages 109–115, Edinburgh, 2003.
- Cheng Niu, Wei Li, Jihong Ding, and Rohini K. Srihari. Orthographic case restoration using supervised learning without manual annotation. *International Journal on Artificial Intelligence Tools*, 13(1):141–156, 2004.
- Rebecca Passonneau. *Instructions for Applying Discourse Reference Annotation for Multiple Applications (DRAMA)*. Department of Computer Science, Columbia University, 1996.
- Massimo Poesio. The MATE/GNOME proposals for anaphoric annotation, revisited. In Michael Strube and Candy Sidner, editors, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 154–162, Cambridge, Massachusetts, USA, April 30 - May 1 2004. Association for Computational Linguistics.
- David Schlangen. Towards finding and fixing fragments: Using ML to identify non-sentential utterances and their antecedents in multi-party dialogue. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 247–254, Ann Arbor, June 2005.
- Rohini K. Srihari, Wei Li, Thomas Cornell, and Cheng Niu. Infoextract: A customizable intermediate level information extraction engine. *Natural Language Engineering*, 12, 2006.
- The PDTB Research Group. The Penn Discourse TreeBank 1.0 annotation manual. Technical Report IRCS-06-01, Institute for Research in Cognitive Science, University of Pennsylvania, March 2006.
- B. Webber. Structure and ostension in the interpretation of discourse deixis. *Language and cognitive processes*, 6(2):107–135, 1991.
- Liang Zhou and Eduard Hovy. Digesting virtual “geek” culture: The summarization of technical internet relay chats. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 298–305, Ann Arbor, Michigan, June 2005.