Comprehension across Application Domains and Languages

Helen M. MENG* and Wai Ching TSUI
Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
hmmeng@se.cuhk.edu.hk, wctsui@se.cuhk.edu.hk
*Tel: (852)-2609-8327   Fax: (852)-2603-5505

ABSTRACT

This work demonstrates that our natural language understanding framework can be applied across application domains and languages with ease. Approaches towards language understanding generally involve much handcrafting, e.g. in writing grammars or annotating corpora, hence portability is a desirable trait in the development of language understanding systems. Our framework for natural language understanding couples semantic tagging with Belief Networks for communicative goal inference, and has delivered promising results in the ATIS (Air Travel Information Systems) domain. This work applies the approach to the *stocks domain*. Furthermore, the approach is extended to *Chinese*, to support a biliteral / trilingual (English with two Chinese dialects) spoken dialog system known as ISIS. We introduce the transformation-based parsing technique for language understanding, and found that it is effective in disambiguating among the various kinds of numeric expressions prevalent in the stocks domain, as well as infer possible semantic categories for out-of-vocabulary words. The nonterminal categories produced by parsing are fed to Belief Networks trained on English or Chinese queries for inferring the user's communicative goal. Our experiments gave a goal identification performance of 94% and 93% for Chinese and English respectively.

Keywords: multilingual, natural language understanding, cross-domains, cross-languages

1. INTRODUCTION

A human-computer interface that enables the user to query for information by means of natural language is considered desirable by many. Natural language understanding technology is usually restricted to domain-specific and language-dependent tasks. However, approaches towards language understanding generally involve much handcrafting, e.g. in writing grammars or annotating corpora. This is a daunting and expensive task and forms a major bottleneck in the development of language understanding systems. Therefore, applicability across application domains and languages is desirable in the development of natural language understanding systems.

Previous work addressing portability issues across domains and languages include: (i) Migration from the English ATIS system to the French MASK (Multimodal-Multimedia Automated Service Kisosk) system (Minker, 1997); (ii) The development of the GALAXY system, which can field spoken and typed questions in four domains (Seneff et al., 1999) – weather information, flight information, air travel planning and navigation, and in multiple languages.

We have previously developed a framework for natural language understanding (NLU) which involves semantic tagging and communicative goal inference using Belief Networks (Meng et al., 1999a). The approach has been applied to the ATIS domain in English, and shown promising results. Our current work demonstrates that this framework is applicable to *both* English and Chinese. This forms the NLU component in a biliteral / trilingual[1] spoken dialog system known as ISIS[2] (Meng et al., 2000). Additionally, the ISIS application domain encompasses real-time stock information inquiries as well as transaction requests. This domain presents new complexities that we have not encountered in the ATIS travel domain previously.

2. OUR NLU FRAMEWORK

Our NLU framework combines semantic tagging with communicative goal inference by Belief Networks (BN). Details can be found in (Meng et al., 1999a). We provide a brief description in the following.

Semantic tagging abstracts the words in a query into a set of semantic concepts, as specified by a set of grammar rules. This renders our approach robust towards extra-grammaticalities and spoken disfluencies, which are ignored at an early stage. The concept tags aim to match the attribute labels for database access, though some tags tend to be syntactic in nature. Sentences are automatically tagged using a transformational procedure. This produces a concept sequence for an input query string, e.g.:

QUERY: *what are the dinner flights from indianapolis to san diego on Wednesday may twelfth*

TAGS: <WHAT> <DUMMY> <MEAL_DESCRIPT> <FLIGHT> <FROM> <CITY_NAME> <TO> <CITY_NAME> <PREP> <DAY_NAME> <MONTH> <DAY>

Training queries that are tagged semantically and then used to train a suite of BNs. A BN is developed for each domain-specific goal, to make a binary decision regarding whether the goal is present or absent in the query. A subset of concepts tags of the user's query form the input to each BN, and the subset for each BN is selected based on the Information Gain measure in

$$P(G_i = 1 | \vec{C}) = \frac{P(\vec{C} | G_i = 1) P(G_i = 1)}{P(\vec{C})} \quad (1)$$

relation to the corresponding goal. Goal inference follows

---

[1] This includes English and two Chinese dialects – Putonghua and Cantonese.

[2] ISIS abbreviates Intelligent Speech for Information Systems.

Baye's Theorem (see Equation 1).

This computation is simplified by assuming independent concepts as are captured in our basic BN topology.[3] The *a posteriori* probability is compared with a pre-set threshold in order to make the binary decision regarding the presence / absence of the goal. The binary decisions across all BNs may be merged by maximizing the *a posteriori* probability for identifying a single goal; or the decisions may be used collectively for the identification of multiple goals. Alternatively, if all BNs vote negative for their goals, the query may be rejected as out-of-domain (OOD).

3. THE ISIS DOMAIN

This work demonstrates that our NLU framework developed for the ATIS domain is also applicable to the ISIS domain. ISIS is a spoken dialog system in the stocks domain. It can handle typed queries in Chinese or English (biliteral), or spoken queries in Putonghua, Cantonese or English (trilingual). Hence our NLU framework is applicable across applications domains (from air travel to stocks), and across languages (from English to Chinese).

To aid the task of domain definition and grammar development, we collected some sample queries in both English and Chinese. We requested that our subjects compose questions that they will ask of a stock broker, e.g. questions on real-time stock quotes, or simulated investor accounts. In this manner, we collected 1407 Chinese queries and 1604 English queries. Examples include:

"Amend my purchase order for HSBC from three to six lots please."

如果和記黃埔升了兩塊錢,請給我賣出兩千股

(Translation: If Hutchison Whampoa rises another two dollars, please sell two thousand shares.)

The stocks domain presents new complexities for natural language understanding. Verbalized numbers abound in the domain-specific queries, and they can refer to stock codes (commonly used in Hong Kong), stock prices, number of lots, number of shares, etc. Consider the query:

"I would like to purchase Cheung Kong at ninety five a share."

Verbalized numbers are parsed to obtain their numeric values, and the numeric expressions are classified into the appropriate semantic category with considerations of *both* left and right contexts. We shall elaborate on this in Section 4.2. Another subtlety in this example is that "a share" really means "per share", and the user is not trying to purchase a single share of Cheung Kong.

Additionally, we have found that in the ISIS domain, contextual information may strongly influence semantic interpretation. Consider the example pair:

"If Hutchison rises another two dollars, sell three lots for me," versus

"If Hutchison drops another two dollars, buy three lots for me."

The two examples are similar in that "two dollars" need to be decoded semantically as an incremental share price. However, it refers to an increment in the former example, and a decrement in the latter.

---

[3] A more complex topology is described in (Meng et al., 1999b).

4. A UNIFIED FRAMEWORK FOR ENGLISH AND CHINESE

This work demonstrates that we have a unified NLU framework that is applicable to both languages. We describe our framework in the following.

4.1 Grammar Development for Semantic Tagging

We hand-designed a set of semantic tags (or concept tags) based on our English sentences. Some syntactic tags are included as well. Had there been more data collected, we believe we could have applied a semi-automatic procedure for acquiring such structures from un-annotated corpora (Siu and Meng, 1999). The set of concept tags forms the pre-terminal categories of our English grammar, and the tags are designed and processed to match the attribute labels for subsequent database access. A real-time data capture component continuously updates a relational database based on a dedicated Reuters satellite feed. As an example, consider the stock "HSBC" with concept tag STOCK_NAME. According to invocations specified in our grammar, this tag automatically invokes a procedure that converts it into 0005.HK (which signifies that it is the stock 0005 from the Stock Exchange of Hong Kong). The new tag 0005.HK matches the RIC[4] code, and can be used directly for database access. Examples of other concept tags include SHARE_PRICE, LOT_NUMBER and SHARE_NUMBER.

Input queries in Chinese are first tokenized into words by means of a greedy algorithm together with a 1100-word lexicon. The lexicon currently covers the 33 constituent stocks in the Hang Seng Index. We maximized the reuse of English concept tags (hand-designed with reference to the English queries) as pre-terminal categories for processing the Chinese sentences we have collected. At this initial stage we are using a single Chinese grammar for both Putonghua and Cantonese queries. For example:

QUERY: 請問你有沒有長實的成交量
(Translation: do you have the trading volume of Cheung Kong)
TOKENIZE WORDS: 請問 你 有沒有 長實 的 成交 量
TAGS: <QUES_TYPE> <STOCK_NAME> <DUMMY> <TRADING_VOL>

Our English and Chinese grammars have 169 and 174 preterminal tags respectively. Of these, 143 are common between the two grammars, achieving about 82% sharing thus far. Examples of language-specific tags include: PREP (for English prepositions); TEENS (for English numbers like "eleven", "twelve", etc.); TWENTY and THIRTY (for the Chinese numbers "廿" and "卅"); and "號" (a number marker for Chinese numbers).

4.2 Transformation-based Parsing

As mentioned previously, the stocks domain presents new complexities for natural language understanding. Much ambiguity exists for numeric expressions, and we need to identify precisely to what they refer, be it share price, lot number, stock code, share number, etc. We can conceive of other complexities – for example, the request *"I'd like to close my*

---

[4] RIC stands for Reuters Instrument Code.

position on Intel" may be either a buy or sell request depending on the investor's holdings.

In order to handle these complexities in the ISIS domain, we applied the technique of transformation-based parsing. Transformation-based parsing (TBP) was proposed by Brill (Brill, 1993) and has been applied to corpora such as the Penn Treebank. TBP has also been shown to achieve better performance than a corpus-derived PCFG (Brill, 2000). In TBP, the grammar consists of a sequence of precedence-ordered transformation rules. To the best of our knowledge, this work is one of the first attempts to apply TBP towards natural language understanding. We found that the context-sensitive transformations offer a powerful solution for handling the ambiguities in our current domain. We illustrate with the following examples:

- **Handling Numeric Expressions**

QUERY: *Sell two more lots of Cheung Kong when it gains another two dollars per share.*

TAGS: <SELL> <NUM_EXP> <MORE> <LOT> <PREP> <STOCK_NAME> <WHEN> <DUMMY> <RISE> <MORE> <SHARE_PRICE>

In this query, "*two dollars*" is an incremental share price relative to the current share price. Our intermediary grammar rules will be able to parse "*two dollars* per share" to be a SHARE_PRICE. However the transformation rule:

RULE: SHARE_PRICE +SHARE_PRICE PREVBIGRAM RISE MORE
which specifies that we should change the concept tag SHARE_PRICE to +SHARE_PRICE (i.e. an incremental share price) if its previous bigram is RISE MORE.

A similar technique can be used to handle the phrase "*two more lots*" in the example query above. Here a pair of precedence-ordered transformation rules are applied:

RULE: NUM_EXP LOTS NEXT1OR2TAG LOT

RULE: LOTS +LOTS NEXTTAG MORE

where the first rule indicates that the numeric expression should be a specified lot number if its following one or two tags is LOT; and the second rule changes the concept tag LOT to +LOT (i.e. the incremental lot number) if its following tag is MORE.

- **Handling Out-of-Vocabulary (OOV) Words**

Language understanding is hampered by words that lie outside of the vocabulary specified in the grammar's terminal categories. However, we have developed a technique that makes use of the context-sensitive transductions to infer a possible concept tag for the OOV. To illustrate with an example, assume that PCCW (Pacific Century CyberWorks Limited) is a stock name which is an OOV.

QUERY: *I'd like to check the news about PCCW over the past two weeks.*

TAGS: <DUMMY> <CHECK> <NEWS> <ABOUT> <OOV> <PREP> <RELATIVE_DATE>

Here our semantic tagger has tagged "PCCW" as <OOV>. However we have specified the transformation:

RULE: OOV STOCK_NAME_OOV PREVBIGRAM NEWS ABOUT

which states that an OOV tag may be inferred as a new stock name if it is preceded by the tag bigram NEWS ABOUT. Hence we have a conjecture regarding the semantic category of the OOV. Within the context of a dialog system, this partial information may contribute towards the generation of an intelligent response such as "Sorry, I do not know about the stock PCCW."

A similar technique is used to handle OOV words in Chinese queries. OOV words are often tokenized as a sequence of mono-character or bi-character units, e.g.

QUERY: 我想查詢榮電國際的最新價格
(Translation: I would like to ask for the latest share price of Global Link.)

Since the stock name is not known by the system, its characters are tagged as OOV. We perform "n-gram grouping" whereby the entire sequence is grouped into a single unit to form "榮電國際", which is in turn given a single OOV tag. Thereafter, transformation rules for OOV tags (similar to the one presented in the English example above) are applied to transform the tag from OOV to STOCK_NAME_OOV.

These transformation rules further modify our set of concept tags, and at this point there approximately 110 concept tags in our English and Chinese grammars respectively. Among only 4 are language-specific tags, which shows a high degree of sharing across languages. Henceforth, all our queries are represented as a sequence of such concept tags.

4.3 Goal Inference with BN

Goal inference with Belief Networks (BN) follows the procedure we have previously used for the ATIS domain (Meng et al., 1999a) We modeled the ISIS domain with 10 domain-specific goals, e.g. quotes, news, chart, purchase_order, etc. Then we went through our English and Chinese queries and annotated each with the most appropriate communicative goal to form our English and Chinese training corpora. A small number of queries are labeled with *multiple* goals, e.g.

QUERY: *please show me the daily chart of HSBC and the closing price of Hang Lung*
GOAL(S): chart, quote

QUERY: 請告訴我匯豐的現價還有給我在一百塊的時候購入二千股
(Translation: Please give me the latest share price of HSBC and buy two thousand shares when it hits a hundred dollars per share.)
GOAL(S): quote, purchase_order

In order to acquire a disjoint test set, we further collected 484 Chinese queries and 532 English queries. These were also annotated with communicative goals. Table 1 summarizes the statistics of our corpora.

	# Chinese queries	# English queries
Training sentences	1407	1604
Single goal sentences (train)	1390	1583
Multiple goal sentences (train)	17	21
Testing sentences	484	543
Single goal sentences (test)	466	532
Multiple goal sentences (test)	6	6
Out-of-domain (test)	12	5

Table 1. Statistics of our training and testing corpora. Out-of-domain queries are not used in training.

While out-of-domain (OOD) queries are not used in training, a few are included in our testing corpora for investigating the capability of rejection. Examples of the OOD queries include:
QUERY: 請告訴我星加坡元兌港元的匯率
QUERY: *Give me the exchange rate for the Singapore Dollar against the Hong Kong Dollar* (Translation of the above)

- **Goal Identification Experiments**

We developed 10 BNs, one for each goal. For each goal, we computed the Information Gain for all goal-concept pairs to select the subset of the N concepts[5] as input to the corresponding BN. Goal inference proceeds as specified in Equation (1). The threshold used for each BN is chosen such that it optimizes the F-measure based on the training data. The BNs with optimized thresholds are used to perform goal identification on our test set queries. In cases where multiple goals are hypothesized for single-goal queries, we penalize the insertion errors. Within-domain queries that are wrongly rejected as OOD, or OOD queries that are wrongly identified as within-domain are all penalized as errors.

Performance on our English test set yields a goal identification accuracy of 92.5%. All the OOD queries were correctly rejected. But only 3 of the 6 multiple-goal queries were correctly identified.

Performance on our Chinese test set yields a goal identification accuracy of 93.7%. 11 of the 12 OOD queries were correctly rejected, and again only 3 of the 6 multiple-goal queries were correctly identified.

The identified goal together with the concepts and their values form the semantic frame that represents the meaning of the query. This is used to formulate a SQL expression for database access.

We have demonstrated that our NLU framework is applicable across languages and domains. The collected data for the ISIS domain is limited at present, but is sufficient for the development of an initial system prototype, with which we can collect more domain-specific data, and bootstrap from there.

- **A Cross-Language Experiment**

We also conducted a side experiment where the BNs trained on Chinese were used for goal identification on the English queries. Both English training and testing sets were used since they are disjoint from the Chinese training corpus. Goal identification accuracy was 82.9%.

Similarly, the BNs trained on English queries were also used for goal identification of the Chinese training and testing sets. Goal identification accuracy was 80.5%. When compared to the monolingual experiments, the main degradation was concentrated on a few goals, and is caused by language-specific elements as well as some degree of over-fitting of the small amount of training data.

We consider this to be an interesting side experiment, as it illustrates that the BN framework can largely apply across languages.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have demonstrated that our NLU framework is applicable across domains and languages. Our framework consists of a semantic tagging procedure followed by communicative goal inference using Belief Networks. This framework was previously used for English in the ATIS (air travel) domain, and is currently extended to handle both English and Chinese for the ISIS (stocks) domain. The stocks domain

presents new complexities for NLU which were not observed in the air travel domain – the prominent ones being the disambiguation of various kinds of numeric expressions, and new words (OOV). We applied a transformation-based parsing technique, which utilizes both the left and right contexts for disambiguating the numeric expressions, and infer a possible category of the OOV word. The concept tags resulting from parsing are fed into BNs previously trained from small sets of collected English and Chinese queries (about 1400 and 1600 for Chinese and English respectively). Experiments with small test sets (about 400 for Chinese and 500 for English) gave goal identification accuracies of 93% for English queries and 94% for Chinese queries. These initial results demonstrate that our NLU framework is applicable to Chinese and the ISIS domain. This NLU component has been integrated into the initial ISIS system prototype. We plan to use the prototype to collect more domain-specific data, and the data can be used for further NLU development and improvement. A side experiment with cross-language configurations gave goal-identification accuracies of 83% (testing English queries on BNs trained on Chinese queries) and 80% (testing Chinese queries on BNs trained on English queries), which suggest that much commonality is captured across the two languages. Future work includes semi-automatic grammar induction (Siu & Meng, 1999), and the use of automatically learned BN topologies (Meng et al., 1999b).

6. ACKNOWLEDGMENTS

7. REFERENCES

Brill, E., "Automatic Grammar Induction and Parsing Tree Free Text: A Transformation-Based Approach," Proceedings of ACL, 1993.

Brill, E., "Automatic Grammar Induction: "Combining, Reducing and Doing Nothing," Proceedings of IWPT, 2000.

Meng, H., W. Lam and C. Wai, "To Believe is to Understand," Proceedings of Eurospeech, 1999a.

Meng, H., W. Lam and K.F.Low, "Learning Belief Networks for Natural Language Understanding," Proceedings of ASRU, 1999b.

Siu, K.C. and H. Meng, "A Semi-Automatic Approach for Acquiring Domain-Specific Semantic Structures," Proceedings of Eurospeech, 1999.

Meng, H., et al., "ISIS: A Multilingual Spoken Dialog System developed with CORBA and KQML Agents," Proceedings of ICSLP, 2000.

Minker, W., "Stochastically-Based Natural Language Understanding Across Tasks and Languages," Proceedings of Eurospeech, 1997.

Seneff, S., et al., "Organization, Communication and Control in the Galaxy-II Conversational System," Proceedings of Eurospeech, 1999.

---

[5] A subset of N=20 concepts is selected. The parameter is chosen based on our previous experience in the ATIS domain. Optimization experiments will be run when more data is available for the ISIS domain.