Understanding questions and finding answers: semantic relation annotation to compute the Expected Answer Type

Volha Petukhova

Spoken Language Systems, Saarland University, Germany
v.petukhova@lsv.uni-saarland.de

Abstract

The paper presents an annotation scheme for semantic relations developed and used for question classification and answer extraction in an interactive dialogue based quiz game. The information that forms the content of this game is concerned with biographical facts of famous people's lives and is often available as unstructured texts on internet, e.g. Wikipedia collection. Questions asked as well as extracted answers, are annotated with dialogue act information (using the ISO 24617-2 scheme) and semantic relations, for which an extensive annotation scheme is developed combining elements from TAC KBP slot filling and TREC QA tasks. Dialogue act information, semantic relations and identified focus words (or word sequences) are used to compute the Expected Answer Type (EAT). Our semantic relation annotation scheme is defined and validated according to ISO criteria for design of a semantic annotation scheme. The obtained results show that the developed tagset fits the data well, and that the proposed approach is promising for other query classification and information extraction applications where structured data, for example, in the form of ontologies or databases, is not available.

Keywords: semantic annotation, annotation scheme design, semantic relations

1. Introduction

According to the ISO Linguistic Annotation Framework (ISO, 2009), the term 'annotation' refers to linguistic information that is added to segments of language data and/or nonverbal communicative behaviour. Semantic annotations have been proven to be useful for various purposes. Annotated data is used for a systematic analysis of a variety of language phenomena and recurring structural patterns. Corpus data annotated with semantic information are also used to train machine learning algorithms for the automatic recognition and prediction of semantic concepts. Finally, semantically annotated data is used to build computer-based services and applications. One of the first steps in obtaining such annotations is the design of a semantic annotation scheme that fits the data well. The International Organization for Standards (ISO) has set up a series of projects for defining standards for the annotation of various types of semantic information, together forming the so-called Semantic Annotation Framework (SemAF). Different parts of SemAF are concerned with (1) time and events; (2) dialogue acts; (3) semantic roles; (4) spatial information; and (5) discourse relations. They define general theoretically and empirically well-founded domain- and language-independent concepts. This presents a good starting point for designing domain-specific schemes, if desired.

In this paper we discuss the design of a domain-specific annotation scheme for semantic relations used for a domain-specific Question Answering (QA) application. In a domain-specific QA, questions are expected about a certain topic; if a question outside that topic is asked, it will not be answered by the system.

The system described here is an interactive guessing game in which players ask questions about attributes of an unknown person in order to guess his/her identity. The player may ask ten questions of various types, and direct questions about the person's name or alias are not allowed. Moreover, the system is a Question Answering Dialogue System (QADS), where answers are not just pieces of extracted text or information chunks, but full-fledged natural language dialogue utterances. The system has all components that any traditional dialogue system has: Automatic Speech Recognition (ASR) and Speech Generation (e.g. TTS) modules, and the Dialogue Engine. The Dialogue Engine, in turn, consists of four components: the interpretation module, the dialogue manager, the answer extraction module and the utterance generation module. The dialogue manager (DM) takes care of overall communication between the user and the system. It gets as input a dialogue act representation from the interpretation module (IM), which it is usually about a question which is uttered by the human player. Questions are classified according to their communicative function (e.g. Propositional, Check, Set and Choice Questions) and semantic content. Semantic content is determined by Expected Answer Type (EAT), e.g. LOCATION as semantic relation, and the focus word, e.g. *study*. To extract the requested information, a taxonomy is designed comprising 59 semantic relations to cover the most important facts in human life, e.g. birth, marriage, career, etc. The extracted information is mapped to the EAT, and both the most relevant answer and a strategy for continuing the dialogue are computed. The DM then passes the system response along for generation, where the DM input is transformed into a dialogue utterance (possibly a multimodal and multifunctional one).

The paper is structured as follows. Section 2 gives an overview of previous approaches to designing semantic relation tagsets for QA applications. Section 3 discusses design criteria for the new semantic relation annotation scheme. Section 4 defines the semantics of the relations and groups them into a hierarchical taxonomy. Section 5 describes the collection of dialogue data and annotations, with indicated reliability of the defined annotation scheme in terms of inter-annotator agreement. In Section 6 classification results using semantic relations in questions and for answer extraction are presented. Section 6 concludes the reported study and outlines future research.

2. Related work

A major breakthrough in QA has been made by (Moldovan et al., 2000) when designing an end-to-end open-domain QA system. This system achieved the best result in the TREC-8 competition[1] with an accuracy of 77.7%. Their system contains the three components: question processing, paragraph indexing and answer processing. First, the question type, question focus, question keyword and expected answer type are specified. There are 9 question classes (e.g. *'what'*, *'who'*, *'how'*) and 20 sub-classes (e.g. *'basic what'*, *'what-who'*, *'what-when'*). Additionally, expected answer type is determined, e.g. *person*, *money*, *organization*, *location*. Finally, a focus word or a sequence of words is identified in the question, which disambiguates it by indicating what the question is looking for (see Moldovan et al., 2000 for an overview of defined classes for 200 of the most frequent TREC-8 questions).

Li and Roth (2002) proposed another question classification scheme, also based on determining the expected answer type. This scheme is a layered hierarchical one having two levels. The first level represents coarse classes like *Date*, *Location*, *Person*, *Time*, *Definition*, *Manner*, *Number*, *Price*, *Title*, *Distance*, *Money*, *Organization*, *Reason* and *Undefined*. The second level has 50 fine-grained classes like *Description*, *Group*, *Individual* and *Title* for the upper-level class of *Human*.

The most recent work comes from the TAC KBP slot filling task (Joe, 2013) aiming to find filler(-s) for each identified empty slot, e.g. for a person (e.g. date_of_birth, age, etc.) and/or for an organization (e.g. member_of, founded_by, etc). Pattern matching, trained classifiers and Freebase[2] are used (Min et al., 2012) and (Roth et al., 2012) to find the best filler. The best system performance achieved in terms of F-score is 37.28% (see Surdeanu, 2013 and Roth et al., 2013).

We see that semantic relations are commonly used to compute an expected answer type. Our task, domain and data differ from the above mentioned approaches in that (1) our domain is closed, (2) the content is mainly unstructured internet articles, and (3) the answers are not just extracted chunks or slot fillers, but rather full dialogue utterances. These aspects cannot be captured by existing annotation approaches. Therefore, we propose a new semantic relation annotation scheme and when developing it we rely on criteria formulated for semantic annotation ISO standards design (see e.g. ISO 24617-2). These criteria support well-founded decisions when designing the conceptual content and structure of the annotation scheme. We discuss the criteria in the next Section.

3. Annotation scheme design criteria

The design of a scheme for annotating primary language data with semantic information is subject to certain methodological requirements, some of which have been made explicit in various studies (Bunt and Romary, 2002; Ide et al., 2003; Bunt and Romary, 2004), and some of which have so far remained implicit. For example, Bunt and Romary (2002) introduce the principle of *semantic adequacy*, which is the requirement that semantic annotations should have a semantics. This is because a semantic annotation is meant to capture something of the meaning of the annotated stretch of source text, but if the annotation does not have a well-defined semantics, then there is no reason why the annotation should capture meaning any better than the source text itself.

A semantic annotation scheme is intended to be applied to language resources, in particular to collections of empirical data. It should therefore contain concepts for dealing with those phenomena which are found in empirical data, allowing good coverage of the phenomena of interest.

Finally, an annotation scheme should be practically useful, i.e. be effectively usable by human annotators and by automatic annotation systems; it should not be restricted in applicability to source texts in a particular language or group of languages; and it should incorporate common concepts of existing annotation schemes where possible.

From these considerations, the following general criteria can be distilled:

- *compatibility*: incorporate common concepts of existing annotation schemes, thus supporting the mapping from existing schemes to the new one, and ensuring the interoperability of the defined scheme.
- *theoretical validity*: every concept defined has a well-defined semantics.
- *empirical validity*: concepts defined in the scheme correspond to phenomena that are observed in corpora.
- *completeness*: concepts defined in the scheme provide a good coverage of the semantic phenomena of interest.
- *distinctiveness*: each concept defined in the scheme is semantically clearly distinct from the other concepts defined.
- and *effective usability*: concepts defined in the scheme are learnable for both humans and machines with acceptable precision.

We will show in this paper that each of these criteria is fulfilled, supporting well-founded decisions when designing the conceptual content and structure of the proposed annotation scheme.

4. Semantic relations

In order to find the answer to a certain question, semantic role information can be used. A semantic role is a relational notion (between an event and its participant) and describes the way a participant plays in an event or state (first defined as such in (Jackendoff, 1972) and (Jackendoff, 1990)), as described mostly by a verb, typically providing answers to questions such as ”who” did ”what” to ”whom,” and ”when,” ”where,” ”why,” and ”how.” Several semantic role annotation schemes have been developed in the past, e.g. FrameNet (ICSI, 2005), PropBank (Palmer et al., 2002), VerbNet (Kipper, 2002) and Lirics (Petukhova and Bunt, 2008).

[1] http://trec.nist.gov/pubs/trec8

[2] http://www.freebase.com/

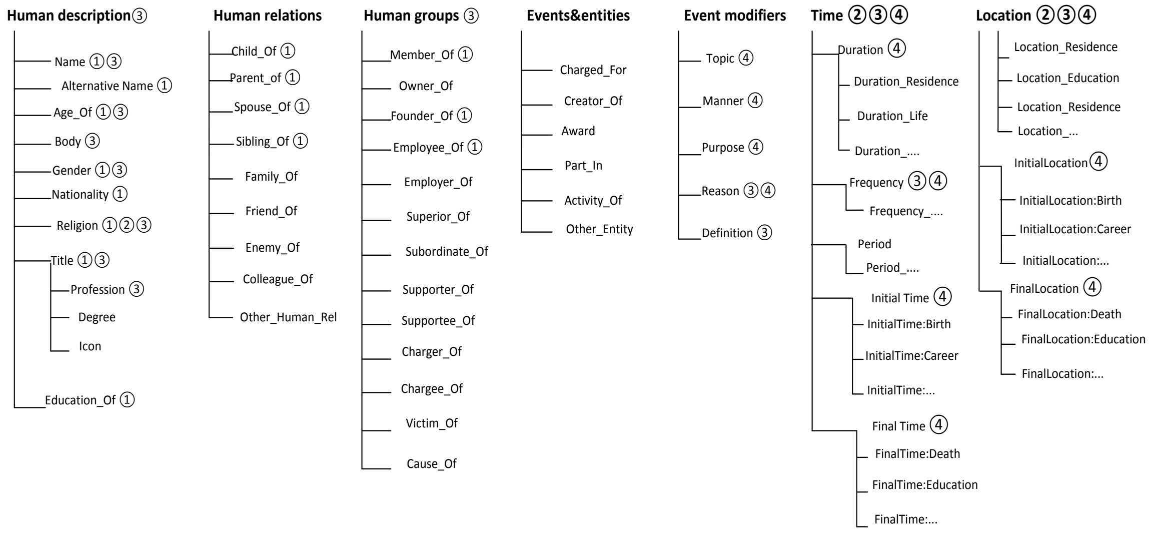

Figure 1: Semantic relations taxonomy. (① means that the relation is also defined in TAC KBP slot filling task; ② in TREC-08 QA task; ③ in TREC 2002 QA task, i.e. annotation scheme proposed by (Li and Roth, 2002); and ④ in LIRICS semantic role set)

Communicative function	%
Propositional Questions	22.4
Set Questions	38.8
Choice Questions	10.4
Check Questions	23.9
Unspecified Question Type	4.5

Table 1: Distribution of information-seeking communicative functions in the annotated data.

Along with semantic roles, relations between participants are also relevant for our domain, e.g. the relation between Agent and Co-Agent (or Partner) involved in a 'work' event may be a COLLEAGUE_OF relation.

To decide on the set of relations to investigate, we analysed available and collected new dialogue data. As a starting point, we analysed recordings of the famous US game 'What's my line?' that are freely available on Youtube (www.youtube.org). However, the latter differs from our scenario: during the TV-show participants may ask only propositional questions with expected 'yes' or 'no' answers;, our game allows any question type from the user. Therefore, we collected data in pilot dialogue experiments, where one participant was acting as a person whose name should be guessed and the other as a game player. 18 dialogues were collected of total duration of 55 minutes comprising 360 system's and user's speaking turns. To evaluate the relation set and to train classifiers, we performed large scale gaming experiments in a Wizard of Oz setting (see Section 4).

Pilot experiments showed that all players tend to ask similar questions about gender, place and time of birth or death, profession, achievements, etc. To capture this information we defined 59 semantic relations. We proposed a multi-layered taxonomy: a high level, coarse annotation comprising 7 classes and a low-level, fine-grained annotation, comprising 52 classes. This includes the HUMAN DESCRIPTION class defined for basic facts about an individual like age, title, nationality, religion, etc.; HUMAN RELATIONS for parent-child and other family relations; HUMAN GROUPS for relations between colleagues, friends, enemies, etc.; EVENTS&NON-HUMAN ENTITIES class for awards, achievements, products of human activities, etc.; EVENT MODIFIERS for specifying manner, purpose, reasons, etc.; the TIME class to capture temporal information like duration, frequency, period, etc.; and the LOCATION class to capture spatial event markers for places where events occur. Some of the second-level classes are broken down into even more specific classes. For example, TITLE has three classes such as PROFESSION for official name(s) of the employment and occupation/job positions; DEGREE for unofficial and official names of obtained degrees and degrees within an organization, e.g. 'highest paid athlete', 'doctor in physics', 'senior leader', etc.; and ICON for unofficial or metaphorical titles that do not refer to an employment or membership position, e.g. 'public figure', 'hero', 'sex symbol', etc. Figure 1 shows the defined hierarchical taxonomy with an indication of what concepts can be found in existing schemes for annotating semantic relations and semantic roles. It should be noted here that the majority of the concepts defined here are domain-specific, i.e. tailored to our quiz game application. The approach could however be adapted for designing comparable annotation schemes for other domains; this has for example been done for the food domain (see Wiegand and Klakow, 2013).

From a semantic point of view, each relation has two arguments and is one of the following types:

- RELATION(Z,?X), where Z is the person in question and X the entity slot to be filled, e.g. CHILD_OF(einstein,?X);
- RELATION(E_1, ?E_2) where E_1 is the event in question and E_2 is the event slot to be filled, e.g. REA-

RELATION	%	RELATION	%	RELATION	%	RELATION	%
ACTIVITY_OF	10.21	LOC_BIRTH	2.34	AGE_OF	3	LOC_DEATH	1.69
AWARD	4.4	LOC_RESIDENCE	1.69	BODY	1.5	MANNER	1.12
CHARGED_FOR	4.21	MEMBER_OF	2.43	CHILD_OF	1.5	NAME	1.87
COLLEAGUE_OF	1.03	NATIONALITY	1.22	CREATOR_OF	6.09	OWNER_OF	1.97
DESCRIPTION	4.12	PARENT_OF	1.31	DURATION	1.31	REASON	1.22
EDUCATION_OF	3.65	RELIGION	2.53	EMPLOYEE_OF	1.59	SIBLING_OF	0.94
ENEMY_OF	1.12	SPOUSE_OF	1.4	FAMILY_OF	1.59	SUPPORTED_BY	0.94
FOUNDER_OF	1.87	TIME	7.96	FRIEND_OF	1.03	TIME_BIRTH	2.06
GENDER	1.69	TIME_DEATH	1.59	LOCATION	4.68	TITLE	11.14

Table 2: Question types in terms of defined semantic relations and their distribution in data (relative frequency in %).

SON(death,?E_2); and

- RELATION(E,?X) where E is the event in question and X the entity slot to be filled, e.g. DURATION(study,?X).

The slots to be filled are categorized primarily based on the type of entities which we seek to extract information about. However, slots are also categorized by the *content* and *quantity* of their fillers.

Slots are labelled as *name*, *value*, or *string* based on the content of their fillers. *Name* slots are required to be filled by the name of a person, organization, or geo-political entity (GPE). *Value* slots are required to be filled by either a numerical value or a date. The numbers and dates in these fillers can be spelled out (December 7, 1941) or written as numbers (42; 12/7/1941). *String* slots are basically a "catch all", meaning that their fillers cannot be neatly classified as names or values.

Slots can be *single-value* or *list-value* based on the number of fillers they can take. While single-value slots can have only a single filler, e.g. date of birth, list-value slots can take multiple fillers as they are likely to have more than one correct answer, e.g. employers.

5. Data collection and annotations

In order to validate the proposed annotation scheme empirically, two types of data are required: (1) dialogue data containing player's questions that are more realistic than youtube games and larger than our pilots; and (2) descriptions containing answers to player's questions about the guessed person. This data is also required to build an end-to-end QADS.

To collect question data we explored different possibilities. There is some question data publicly available, e.g. approximately 5500 questions are provided by the University Illinois[3] annotated according to the scheme defined in (Li and Roth, 2002). However, not all of this data can be used for our scenario. We filtered out about 400 questions for our purposes. Since this dataset is obviously too small, we generated questions automatically using the tool provided by (Heilman and Smith, 2009) from the selected Wikipedia articles and filtered them out manually. Out of the generated 3000 questions relevant ones were selected: grammatically broken questions were fixed and repetitions deleted. Additionally, synonyms from WordNet[4] were used to generate different variations of questions for the same class. Questions collected in pilot experiments were added to this set as well. The final question set consists of 1069 questions. These questions are annotated with (1) communicative function type according to ISO 24617-2; (2) with semantic relations as defined in Section 3; and (3) with question focus word or word sequence. Table 1 provides an overview of the types of information-seeking communicative functions in the collected data and those relative frequencies.

Table 2 illustrates the distribution of question types based on the EAT's semantic relation.

A focus word or word sequence describes the main event in a question, usually specified by a verb or eventive noun. The focus word (sequence) is extracted from the question to compute the EAT and formulate the query. For example,

(1) Question: When was his first album released?
Assigned semantic relation: TIME
Focus word sequence: first album released
EAT: TIME_release(first_album)
Query:
TIME_release(first_album) :: (E, ?X) :: QUALITY(VALUE) :: QUANTITY(SINGLE)

The question set is currently enriched with questions from large scale Wizard of Oz experiments. The data collection procedure was similar to that of pilots. A Wizard (English native speaker) simulated the system's behaviour and the other participant played the game. 21 unique subjects, undergraduates of age between 19 and 25, who are expected to be related to our ultimate target audience, participated in these experiments. 338 dialogues were collected of a total duration of 16 hours comprising about 6.000 speaking turns. An example from this dialogue collection can be found in the Appendix.

Answers were retrieved from 100 selected Wikipedia articles in English containing 1616 sentences (16 words/sentence on average), 30.590 tokens (5.817 unique tokens). Descriptions are annotated using complex labels consisting of an IOB-prefix (**I**nside, **O**utside, and **B**eginning), since we aim to learn the exact answer boundaries, and semantic relation tag, the same as used for classifying questions. We mainly focus on labeling nouns and noun phrases. For example:

[3] http://cogcomp.cs.illinois.edu/page/resources/data

[4] urlhttp://wordnet.princeton.edu/

RELATION	%	RELATION	%	RELATION	%	RELATION	%	RELATION	%
ACCOMPLISHMENT	4.0	DURATION	1.8	LOC_DEATH	0.8	PART_IN	3.6	TIME	14.6
AGE_OF	2.1	EDUCATION_OF	4.2	LOC_RESIDENCE	3.2	RELIGION	0.7	TIME_BIRTH	2.8
AWARD	2.5	EMPLOYEE_OF	2.2	MEMBER_OF	1.8	SIBLING_OF	2.3	TIME_DEATH	1.0
CHILD_OF	3.6	FOUNDER_OF	1.2	NATIONALITY	3.1	SPOUSE_OF	1.9	TITLE	14.2
COLLEAGUE_OF	1.7	LOC	5.6	OWNER_OF	1.1	SUBORDINATE_OF	1.3		
CREATOR_OF	8.5	LOC_BIRTH	5.0	PARENT_OF	3.7	SUPPORTEE_OF	1.1		

Table 3: Answer types in terms of defined semantic relations and their distribution in data (relative frequency in %)

(2) *Gates graduated from* ***Lakeside School*** *in 1973.*

The word *Lakeside* in (2) is labeled as the beginning of an EDUCATION_OF relation (B-EDUCATION_OF), and *school* is marked as inside of the label (I-EDUCATION_OF). Table 3 illustrates the distribution of answer types based on the identified semantic relation.

Since the boundaries between semantic classes are not always clear, we allowed multiple class labels to be assigned to one entity. For example:

(3) *Living in Johannesburg, he became involved in anti-colonial politics, joining the ANC and becoming a founding member of its* ***Youth League****.*

Here, *Youth League* is founded by a person (FOUNDER_OF relation), but the person is also a member of the *Youth League*. There are also some overlapping segments detected as in example (4):

(4) *He served as* ***the commander-in-chief of the Continental Army*** *during the American Revolutionary War.*

The entity *commander-in-chief of the Continental Army* in (4) is marked as TITLE, while *the Continental Army* is recognized as MEMBER_OF. Both of these relations are correct, since if a person leads an army he/she is also a member of it.

To assess the reliability of the defined tagset, the inter-annotator agreement was measured in terms of the standard Kappa statistic (Cohen, 1960). For this, 10 randomly selected descriptions and all 1069 questions were annotated by two trained annotators. The obtained *kappa* scores were interpreted as annotators having reached good agreement (averaged for all labels, kappa = .76).

6. Semantic relation classification and learnability

To investigate the learnability of the relations we defined in a data-oriented way and to evaluate the semantic relation set, we performed a number of classification experiments. Moreover, we partition the training sets in such a way that we can assess relation learnability by plotting learning curves for each relation given an increasing amount of training data.

Classifiers used were statistical ones, namely, Conditional Random Fields (CRF) (Lafferty et al., 2001) and Support Vector Machines (SVM) (Joachims et al., 2009).[5].

The selected feature set includes **word & lemma tokens**; **n-grams** and **skip n-grams** for both tokens and their lemmas; **POS** tags from the Stanford POS tagger (Toutanova et al., 2003); **NER** tags from three different NER tools: Stanford NER (Finkel et al., 2005), Illinois NER (Ratinov and Roth, 2009), and Saarland NER (Chrupala and Klakow, 2010); **chunking** using OpenNLP [7] to determine NP boundaries; **key word** to determine the best sentence candidate for a particular relation, e.g. *marry, married, marriage, husband, wife, widow, spouse* for the SPOUSE_OF relation.

System	Precision	Recall	F-score	Accuracy
Baseline	76.87	73.79	73.72	97.38
System 1	80.18	77.71	78.05	97.89

Table 4: Question classification results

To assess the system performance standard evaluation metrics are used, precision (P), recall (R) and F-score (F1). In particular, precision is important, since it is worse for the system to provide a wrong answer than not to provide any answer at all, e.g. to say it cannot answer a question.[8] It should be noted that for answer extraction sequential classifiers were trained and their predictions were considered as correct iff both the IOB-prefix and the relation tag fully correspond to those in the referenced annotation.

6.1. Question classification

In the 10-fold cross-validation classification experiments, classifiers were trained and evaluated in two different settings: (1) *Baseline*, where classification is based solely on the bag-of-words features; (2) and *System 1*: best system performance after trying different sets of features and selection mechanisms, namely, on bag-of-words plus bigrams generated from bag-of-lemmas. Table 4 presents the classification results.

It may be observed that System 1 clearly outperforms the baseline. The results are also better than those of the state-of-art systems on this task. To compare, the system reported in (Dell and Wee Sun, 2003) using SVM reached 80.2% accuracy (using bag-of-words) and 79.2% (using bag-of-ngrams) for the 50 question classes defined in (Li and Roth, 2002) and on their data. The reported in (Huang et al., 2008) the accuracies of SVM and Maximum Entropy (ME) classifiers were 89.2% and 89.0% respectively on the data and taxonomy of (Li and Roth, 2002). The best performance in terms of accuracy reported by Li and Roth (2006)

[5] We used two CRF implementations from CRF++[6]

[7] http://opennlp.apache.org/

[8] Each WoZ experiment participant filled in a questionnaire, where among other things they indicated that 'not-providing' an answer was entertaining; giving wrong information, by contrast, was experienced as annoying.

	Baseline			System 1			System 2		
	P	R	F1	P	R	F1	P	R	F1
CRF ++	0.56	0.34	0.42	0.68	0.52	0.59	0.82	0.55	0.66
SVM-HMM	0.59	0.28	0.38	0.53	0.51	0.52	0.72	0.47	0.57
Pattern*	-	-	-	-	-	-	0.74	0.62	0.67

Table 6: Overall system performance. *) applied only to 12 most frequently occurring relations

of the tagset was 89.3% using the SNoW learning architecture for a hierarchical classifier .

The performance of the classifiers (System 1 setting) on each relation in isolation has also been assessed. Table 5 presents the obtained results.

Our classifiers achieved reasonably high accuracy in detecting all relations. In terms of F-score, three relations were rather problematic, namely OWNER_OF, DESCRIPTION and SUPPORTEE_OF. For the latter, the number of training instances was rather low as we will show in our learnability experiments (see Section 5.3). For the first one, we have concluded that this relation requires a more clear definition to make better distinctions with other classes, e.g. it is often confused with CREATOR_OF and FOUNDER_OF. Similarly, the DESCRIPTION relation has a rather vague definition and tends to be applied for many unclassifiable instances. We introduce two relations instead: DEFINITION and TOPIC (see Figure 1).

6.2. Answer extraction

In the 5-fold cross-validation classification experiments, classifiers were trained and evaluated in three different settings: (1) *Baseline* obtained when training classifiers on word token features only; (2) *System 1* where classification is based on automatically derived features such as n-grams for tokens and lemmas (trigrams), POS, NER tags and chunking; joint classification on all relations; (3) and *System 2*: pattern matching and classification on the same features as System 1 applied for each relation separately.

Both CRF++ and SVM-HMM classifiers in System 1 and 2 settings show gains over the baseline systems. To appreciate how good statistical classifiers generally are on relation recognition for answer extraction, consider the performance of distant supervision SVM[9] with precision of 53.3, recall of 21.8 and F-score of 30.9 (Roth et al., 2013) on the TAC KBP relations. However, we emphasize that our task, relation set, application and data are different from those of TAC KBP.

As can be observed from Table 6, the CRF++ classifier achieves the best results in terms of precision and F-score. Although the running time was not measured, the classification runs faster than the SVM-HMM. System 2 outperforms System 1 (6-11% increase in F-score). When training on each relation in isolation, feature weights can be adjusted more efficiently, while not affecting other classifiers' performances.

More detailed results from CRF++ on each semantic relation classification can be seen in Table 7.

Relation	P	R	F1	Relation	P	R	F1
ACCOMPLISHMENT	0.73	0.44	0.55	NATIONALITY	0.92	0.73	0.81
AGE_OF	0.95	0.76	0.84	OWNER_OF	0.76	0.40	0.48
AWARD	0.80	0.62	0.70	PARENT_OF	0.79	0.54	0.63
CHILD_OF	0.74	0.58	0.65	PART_IN	0.25	0.05	0.08
COLLEAGUE_OF	0.78	0.32	0.43	RELIGION	0.60	0.16	0.24
CREATOR_OF	0.64	0.17	0.26	SIBLING_OF	0.92	0.69	0.78
DURATION	0.97	0.64	0.76	SPOUSE_OF	0.76	0.42	0.52
EDUCATION_OF	0.84	0.65	0.72	SUBORDINATE_OF	0.81	0.19	0.31
EMPLOYEE_OF	0.77	0.19	0.28	SUPPORTEE_OF	1.00	0.40	0.54
FOUNDER_OF	0.65	0.26	0.36	MEMBER_OF	0.65	0.14	0.21
LOC	0.77	0.33	0.45	TIME	0.90	0.83	0.86
LOC_BIRTH	0.94	0.84	0.89	TIME_BIRTH	0.92	0.89	0.90
LOC_DEATH	0.90	0.55	0.67	TIME_DEATH	0.94	0.79	0.86
LOC_RESIDENCE	0.86	0.55	0.66	TITLE	0.84	0.66	0.74

Table 7: CRF++ performance on System 2.

6.3. Learnability

The outcome from the learnability experiments is presented in Figure 2. From these graphs, we can clearly observe that larger training data positively correlates with higher F-score. The SUPPORTEE_OF is the most sensitive relation to the amount of training data, followed by LOC_DEATH and SUBORDINATE_OF.

7. Discussion and conclusions

We propose an annotation scheme for question classification and answer extraction from unstructured textual data based on determining semantic relations between entities. Semantic relation information together with the focus words (or word sequences) is used to compute the Expected Answer Type. Our results show that the relations that we have defined help the system to understand user's questions and to capture the information, which needs to be extracted from the data. The proposed scheme fits the data and is reliable, as evidenced by good inter-annotator agreement. Semantic relations can be learned successfully in a data-oriented way. We found the ISO semantic annotation scheme design criteria very useful. Following them supported our decisions when defining concepts and the structure of the scheme. The proposed approach is promising for other query classification and information extraction tasks for domain-specific applications.

There is a lot of room for further research and development, and the annotation scheme is far from perfect. For instance, observed inter-annotator agreement and classification results indicate that some relations need to be re-defined. We will test how generic the proposed approach is by testing it on the TAC and TREC datasets. Moreover, since some relations, in particular of RELATION(E$_1$, ?E$_2$) and RELATION(E,?X) types, are very close to semantic roles, there is a need to analyse semantic role sets (e.g. ISO semantic roles (Bunt and Palmer, 2013)) and study the possible overlaps.

From the QADS development point of view, we need to evaluate the system in real settings. For this, the ASR is currently retrained, i.e. generic language and acoustic models are adapted to our game scenario. For now, all classification experiments were run on data transcribed by a human. It is a semi-automatic process, when the ASR output has been corrected. The real system, however, needs to operate on ASR output lattices (list of hypotheses for each token with

[9] Distant supervision method is used when no or little labeled data is available, see (Mintz et al., 2009).

Relation	P	R	F1	Accuracy (in %)	Relatio	P	R	F1	Accuracy (in %)
ACTIVITY_OF	0.61	0.72	0.67	92.56	AGE_OF	1.00	0.93	0.96	99.78
AWARD	0.83	0.85	0.84	98.59	BODY	0.54	0.59	0.57	98.64
CHARGED_FOR	0.96	0.87	0.91	99.27	CHILD_OF	0.85	0.76	0.81	99.45
COLLEAGUE_OF	0.63	0.65	0.64	99.25	CREATOR_OF	0.73	0.69	0.71	96.58
DESCRIPTION	0.32	0.42	0.36	93.86	DURATION	0.93	0.99	0.96	99.90
EDUCATION_OF	0.91	0.79	0.85	98.97	EMPLOYEE_OF	0.91	0.75	0.83	99.49
ENEMY_OF	0.81	0.56	0.66	99.35	FAMILY_OF	0.45	0.88	0.59	98.07
FOUNDER_OF	0.85	0.66	0.74	99.14	FRIEND_OF	1.00	0.72	0.84	99.71
GENDER	1.00	0.97	0.99	99.95	LOCATION	0.78	0.91	0.84	98.38
LOC_BIRTH	0.99	0.92	0.95	99.79	LOC_DEATH	0.80	0.89	0.84	99.44
LOC_RESIDENCE	0.93	0.71	0.81	99.42	MANNER	1.00	0.92	0.96	99.91
MEMBER_OF	0.92	0.67	0.77	99.04	NAME	0.95	0.91	0.93	99.73
NATIONALITY	0.97	0.48	0.64	99.34	OWNER_OF	0.42	0.22	0.29	97.86
PARENT_OF	0.74	0.91	0.82	99.46	REASON	1.00	0.61	0.76	99.52
RELIGION	0.99	0.74	0.85	99.34	SIBLING_OF	0.98	0.80	0.88	99.80
SPOUSE_OF	0.78	0.59	0.67	99.19	SUPPORTEE_OF	0.69	0.20	0.31	99.17
TIME	0.94	0.95	0.95	99.16	TIME_BIRTH	0.95	0.85	0.90	99.61
TIME_DEATH	1.00	0.71	0.83	99.53	TITLE	0.73	0.89	0.80	95.01

Table 5: Question classification results for each relation in isolation.(*presented in alphabetic order)

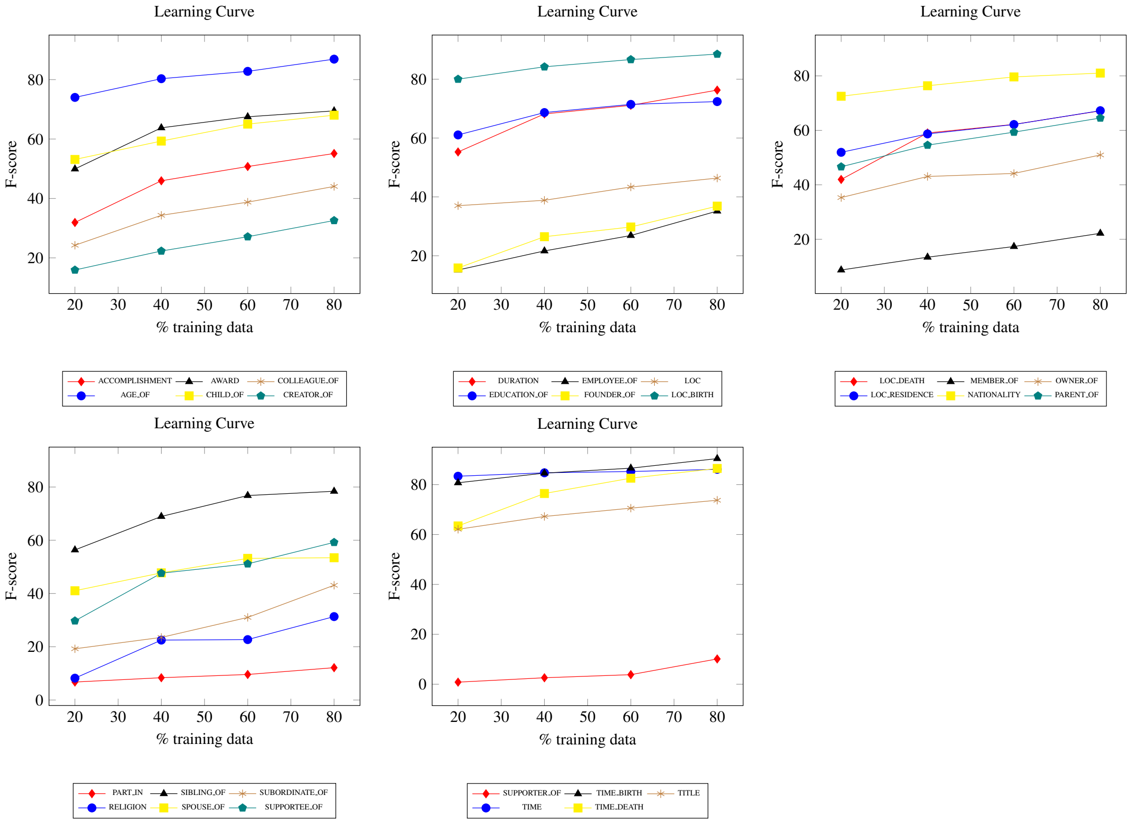

Figure 2: Learning curves for the defined relations

the recognizer's confidence scores). Therefore, in the nearest future we will test the question classifiers performance on the actual ASR output.

8. Acknowledgments

The research reported in this paper was carried out within the DBOX Eureka project under number E! 7152. The author is particularly thankful to her Master students Desmond Darma Putra and Humayun Faiz who performed machine-learning classification experiments.

9. References

H. Bunt and M. Palmer. 2013. Conceptual and representational choices in defining an iso standard for semantic role annotation. In *Proceedings Ninth Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-9)*, Potsdam.

H. Bunt and L. Romary. 2002. Towards multimodal semantic representation. In *Proceedings of LREC 2002 Workshop on International Standards of Terminology and Language Resources Management*, pages 54–60, Las Palmas, Spain.

H. Bunt and L. Romary. 2004. Standardization in multimodal content representation: Some methodological issues. In *Proceedings of LREC 2004*, pages 2219–2222, Lisbon, Portugal.

G. Chrupala and D. Klakow. 2010. A named entity labeler for german: Exploiting wikipedia and distributional clusters. In *Proceedings of LREC'10*, Valletta, Malta. European Language Resources Association (ELRA).

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, 20:37–46.

Z. Dell and L. Wee Sun. 2003. Question classification using support vector machines. In *Proceedings of SIGIR*, pages 26–32, Toronto, Canada.

J. Finkel, T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL '05*, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Heilman and N. Smith. 2009. Question generation via overgenerating transformations and ranking. Language Technologies Institute, Carnegie Mellon University Technical Report CMU-LTI-09-013.

Z. Huang, M. Thint, and Z. Qin. 2008. Question classification using head words and their hypernyms. In *Proceedings of EMNLP*, pages 927–936.

ICSI. 2005. Framenet. Available at `http://framenet.icsi.berkeley.edu`.

N. Ide, L. Romary, and E. de la Clergerie. 2003. International standard for a linguistic annotation framework. In *Proceedings of HLT-NAACL Workshop on The Software Engineering and Architecture of Language Technology*, Edmunton.

ISO. 2009. *ISO 24612:2009 Language resource management: Linguistic annotation framework (LAF)*. ISO, Geneva.

ISO. 2012. *Language resource management – Semantic annotation framework – Part 2: Dialogue acts. ISO 24617-2*. ISO Central Secretariat, Geneva.

R.S. Jackendoff. 1972. *Semantic interpretation in generative grammar*. MIT Press.

R.S. Jackendoff. 1990. *Semantic structures*. MIT Press.

T. Joachims, T. Finley, and C.-N. Yu. 2009. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59.

E. Joe. 2013. Tac kbp 2013 slot descriptions.

K. Kipper. 2002. Verbnet: A class-based verb lexicon. Available at `http://verbs.colorado.edu/~mpalmer/projects/verbnet.html`.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML '01*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

X. Li and D. Roth. 2002. Learning question classifiers. In *Proceedings of the COLING '02*, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

X. Li and D. Roth. 2006. Learning question classifiers: the role of semantic information. *Natural Language Engineering*, pages 229–249.

B. Min, X. Li, R. Grishman, and S. Ang. 2012. New york university 2012 system for kbp slot filling. In *Proceedings of the 5th Text Analysis Conference (TAC 2012)*.

M. Mintz, R. Bills, S.and Snow, and Jurafsky D. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL/IJCNLP)*, page 10031011.

D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Girju, R. Goodrum, and V. Rus. 2000. The structure and performance of an open-domain question answering system. In *Proceedings of ACL '00*, pages 563–570, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Palmer, D. Gildea, and P. Kingsbury. 2002. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

V. Petukhova and H. Bunt. 2008. Lirics semantic role annotation: Design and evaluation of a set of data categories. In *Proceedings of the sixth international conference on language resources and evaluation (LREC 2008)*. Paris: ELRA.

L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of CoNLL '09*, pages 147–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

B. Roth, G. Chrupala, M. Wiegand, M. Singh, and D. Klakow. 2012. Saarland university spoken language systems at the slot filling task of tac kbp 2012. In *Proceedings of the 5th Text Analysis Conference (TAC 2012)*, Gaithersburg, Maryland, USA.

B. Roth, T. Barth, M. Wiegand, M. Singh, and D. Klakow. 2013. Effective slot filling based on shallow distant supervision methods. In *TAC KBP 2013 Workshop*, Gaithersburg, Maryland USA. National Institute of Standards and Technology.

M. Surdeanu. 2013. Overview of the tac2013 knowledge base population evaluation: English slot filling and temporal slot filling. In *TAC KBP 2013 Workshop*, Gaithersburg, Maryland USA. National Institute of Standards and Technology.

K. Toutanova, D. Klein, C. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of NAACL '03*, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Wiegand and D. Klakow. 2013. Towards the detection of reliable food-health relationships. In *Proceedings of the NAACL-Workshop on Language Analysis in Social Media (NAACL-LASM)*.

Appendix: dialogue example

S: Hello
P: Hello
S: Good afternoon almost evening
S: What is your name
P: My name is James
S: Hello James it's nice to meet you
P: Nice to meet you
S: How are you doing today?
P: Good, thank you
S: Alright
S: Today we are going to play a game and here are the rules
S: I'm a very famous person and you need to guess my name you can ask whatever questions you want of me except for my name directly
S: You have at most ten questions and then you get to guess my name exactly once
S: So you can ask whatever questions you want but then if you want to guess my name you only get one try
S: If you get my name correct you win if you get my name incorrect or choose to pass then you lose and then we'll move on to the next round
S: Do you understand and are comfortable with the rules?
P: Yeah yeah
P: So the name is kind of a famous person
P: Okay
P: I'm not sure how good am I in this area
S: Yes
S: I am a famous person and I am male
P: Okay okay good
S: Alright
S: And what is your first question?
P: What is the first question
P: What do you do?
S: I am a leader
P: A leader
P: What is your nationality?
S: I am American
P: Are you alive?
S: I am not alive
P: Are you leading a company?
S: I am not leading a company
P: okay
P: You're not a company leader
P: When are you born?
S: I was born on February twenty second seventeen thirty two
P: Seventeen thirty two
P: Ok
P: Eehm
P: Are a politician?
S: I am a politician
P: Okay
P: So then it is not my area but I will try to guess
P: When were you in the government?
S: Uhm
S: Let's see
S: I retired from the presidency in seventeen ninety seven
P: Ninety seven
P: George Washington
S: Is that your final guess?
P: Yes, Washington
S: Very good, excellent job!
S: Congratulations!