How hard is this query? Measuring the Semantic Complexity of Schema-agnostic Queries

André Freitas[1], Juliano Efson Sales[2], Siegfried Handschuh[1], Edward Curry[2]
[1]Department of Computer Science and Mathematics
University of Passau
[2]Insight Centre for Data Analytics
National University of Ireland, Galway
[1]firstname.lastname@uni-passau.de
[2]firstname.lastname@insight-centre.org

March 25, 2015

Abstract

The growing size, heterogeneity and complexity of databases demand the creation of strategies to facilitate users and systems to consume data. Ideally, query mechanisms should be *schema-agnostic*, i.e. they should be able to match user queries in their own vocabulary and syntax to the data, abstracting data consumers from the representation of the data. This work provides an information-theoretical framework to evaluate the semantic complexity involved in the query-database communication, under a schema-agnostic query scenario. Different entropy measures are introduced to quantify the semantic phenomena involved in the user-database communication, including structural complexity, ambiguity, synonymy and vagueness. The entropy measures are validated using natural language queries over Semantic Web databases. The analysis of the semantic complexity is used to improve the understanding of the core semantic dimensions present at the query-data matching process, allowing the improvement of the design of schema-agnostic query mechanisms and defining measures which can be used to assess the semantic uncertainty or difficulty behind a schema-agnostic querying task.

Semantic Complexity, Entropy, Schema-agnostic Queries, Database Queries, Databases

1 Introduction

The growing data availability on Big Data environments demands the creation of strategies to facilitate the interaction between data consumers and databases. As the number of available data sources grows and schemas increase in size and complexity, the effort associated with matching an *information need* to a database schema, intrinsic to the creation of structured queries such as SPARQL and SQL, becomes prohibitive. Ideally, data consumers, being them humans or intelligent agents, should be able to be abstracted from the representation of the data by using a *schema-agnostic query mechanism* [6].

However, structured queries are still the primary way to interact with databases. Despite the evolution of natural language interfaces (NLIs), and the empirical evaluation behind different NLI approaches, relatively little attention is given to the analysis of the semantic phenomena behind the user-database communication (UDC). The construction of semantic models for databases brings the potential of improving UDC and the design of more principled schema-agnostic query mechanisms.

In this work information theoretic models are used to define measures of *semantic complexity* for *schema-agnostic queries*. The measures of semantic complexity are used to quantify the role of core semantic phenomena such as *ambiguity*, *synonymy* and *matching complexity* in the *semantic interpretation*

Proceedings of the 11th International Conference on Computational Semantics, pages 294–304,
London, UK, April 15-17 2015.

of schema-agnostic queries. The contributions of this paper are: (i) to provide a principled and comprehensive analysis of existing semantic measures of semantic complexity in the UDC context, (ii) to validate these measures over a realistic query scenario based on natural language queries over large-schema RDF graph datasets, (iii) to introduce novel semantic complexity measures based on distributional semantic models and (iv) to use the semantic complexity models to support the design of schema-agnostic queries.

This paper is organized as follows: section 2 introduces schema-agnostic queries; section 3 introduces the concept of semantic complexity and entropy; section 4 describes the schema-agnostic queries and the associated semantic entropy model and measures; section 5 validates the model and discusses design principles derived from the entropy measures which can be used on schema-agnostic query mechanisms; section 6 describes conclusions and future works.

2 Mapping Schema-agnostic Queries

Schema-agnostic queries are queries which assume that users do not know the terminology and the structural relations inside a dataset while expressing their information needs [6, 5]. Since the query information can be represented in the database using different terms and relations, schema-agnostic queries are intrinsically associated with a *semantic matching* and *interpretation* model. Schema-agnostic queries can follow a natural language, keyword or a structured query syntax.

In the Information Retrieval space, different works evaluated the query performance by providing predictors based on language models applied in the estimation of vagueness and ambiguity (clarity score in [3]), and by improving query performance using selective pruning [15]. Sullivan [14] uses effectiveness measures to classify 50 question narratives over unstructured text as easy or hard.

Previous works have investigated the formal conditions for mapping a natural language query to a database. The work of Popescu et al. [12] provides a formal description of *natural language interfaces to databases*, concentrating on the definition of the concept of *semantic tractability*. Essentially, the concept of *semantic tractability* provides a description of soundness and completeness conditions for mapping natural language queries to database elements. Comparatively, this work focuses on evaluating query performance predictors for schema-agnostic queries on structured data, targeting addressing schema-agnostic queries over heterogeneous databases.

3 Semantic Complexity & Entropy

The concept of *entropy* in information theory is defined as a measure of uncertainty or surprise associated with a random variable. The random variable represents possibilities over the possible *states* or *configurations* that a specific symbolic system can be in, where the entropy is directly proportional to the number of states.

In order to transport the concept of entropy to the UDC problem, four symbolic sets are introduced: (i) a *word set* W, which expresses the set of words used to describe the domain of discourse shared by the user and database, (ii) a *word sense set* WS, which describes the possible senses associated with the words, (iii) a *proposition set* S, to describe the possible (syntactically valid) compositions of words senses and (iv) a *concept set* C, to describe the set of concepts associated with the possible interpretation for all the compositions. The unambiguous *semantic interpretation* of a query $I(q)$ or database statement $I(s)$ is a concept c_i in the concept domain. Figure 1 depicts the relationship between the sets in the query/database interpretation process. Ambiguity, vagueness and synonymy are defined as mappings patterns between the four sets.

It is possible to define a set M for the semantically valid mappings between W and C under a specific query database matching $m_\Sigma(Q, G)$ for a specific semantic model Σ. The semantic entropy associated with the query-DB matching is proportional to the cardinality of M.

In the context of schema-agnostic queries, the concept of entropy can be interpreted under four main perspectives:

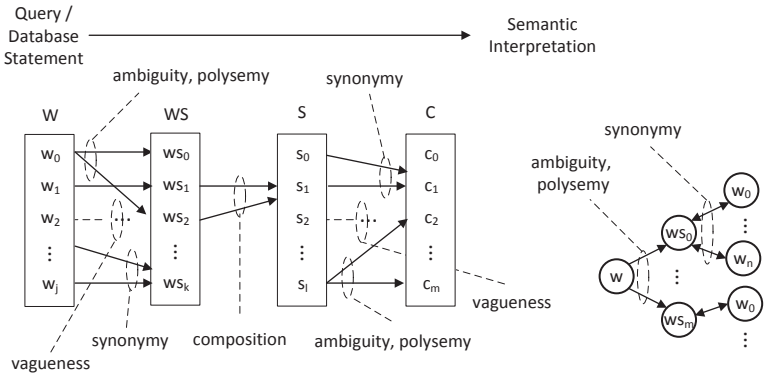

Figure 1: Mapping from words in a query to meaning (w_j) to word sense (ws_k), syntactic composition (s_l) and the associated concept (c_m) for the statement.

- (i) **structural/conceptual complexity:** Databases which express a large number of concepts have larger semantic entropy values. The number of interpretations is usually correlated to the number of distinct entities in the database and the number of possible compositions between them (propositions).
- (ii) **level of ambiguity:** Words/propositions can convey different meanings. The degree of ambiguity (number of possible interpretations) varies for different words and propositions. Depending on the domain of discourse and on the selection of the words, queries and databases can have different levels of associated ambiguity.
- (iii) **vocabulary gap/indeterminacy/vagueness:** The interpretation of a query or of a database statement is dependent on the ability of the data consumer (receiver) to interpret the expressed information. Query and databases may not be expressed in the same vocabulary (synonymy phenomenon) or in the same abstraction-level. Additionally, query and data may not be mapped with the contextual information available in the query or in the database. Indeterminacy/vagueness are semantic phenomena where words, entities or propositions fail to map to the exact meaning intended by the transmitter.
- (iv) **novelty:** Semantic entropy is usually associated with the degree of novelty/informativeness/-surprise associated with the communication process. The more informative the result returned by a query in relation to the specific background knowledge of the query issuer, the larger the entropy value. This dimension is not the focus of this work.

The process of mapping a schema-agnostic query Q to a database associated interpretation $I_G(Q)$ depends on the semantic entropy associated with each entropy dimension and involves coping with the semantic phenomena of structural complexity, term ambiguity, structural ambiguity, vagueness and synonymy. The next section introduces *semantic entropy measures* for each of these dimensions. In the definition of the entropy measures, a practical perspective was adopted (which focuses on the computation of these measures instead of a purely formal model) where the definition of approximate measures take place wherever the application of the complete model is not viable or practical.

4 Semantic Entropy Measures

A generic interpretation process for a schema-agnostic query Q can be defined as a set of steps which map a sequence of words $<w_0, w_1, ..., w_n>$into a set of possible database interpretations $I_G(Q)$. It is assumed that both query and database terminologies are defined under the same language L and that database entities are described using natural language labels. The generic process of interpreting the query can be summarized into the following steps with a set of associated entropy measures:

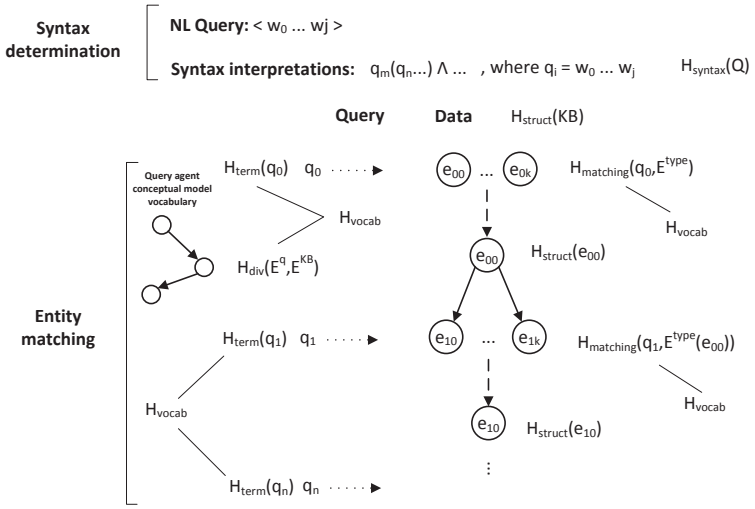

Figure 2: Generic steps for the query processing and associated entropy measures for each step.

- **Syntactic matching:** Consists in the possible interpretations for the syntactic structure of the query under the database syntax. This step consists in the segmentation of the query Q into a set of terms $<q_0, q_1, ..., q_n>$. The entropy H_{syntax} expresses the syntactic uncertainty/ambiguity in the determination of the syntactic mapping.
- **Vocabulary matching:** Consists in the matching/alignment between query entities and database entities, once a syntactic structure was defined. The entropy H_{vocab} is the uncertainty/ambiguity associated with the matching between query entity candidates and database entities.

Figure 2 depicts the steps in the query interpretation process and the associated entropies, while Figure 3 depicts an example for a specific query example. In this section, to maximize generalizability a logical (constant, predicate) terminology is used to express database statements and queries. In the evaluation section the model is specialized into the RDF/SPARQL model.

4.1 Measures of Semantic Entropy

4.1.1 Syntactic Entropy (H_{syntax})

The syntactic entropy of a query is defined by the possible syntactic configurations in which a query can be interpreted under the database syntax. Figure 2 and Figure 3(2) depicts H_{syntax} within the query interpretation model. The syntactic interpretation of a query Q is a tuple $T = <C, \Pi, R, L, Op>$, where C and Π are the set of constants and predicates in the database, $R \rightarrow \Pi \times C \times \cdots$ is the ordered set of syntactic n-ary associations between C and Π, L is the set of logical operators $\wedge, \vee$ and Op a set of functional operators.

The syntactic entropy is given as a function of the probability of the syntactic interpretation of a query. Let Syn be the *lexical categories* and *constituent categories* associated with the set of query words w_i and terms q_i. Let DM be the data model categories (e.g. C, Π, R, L, Op) in which the set of Syn categories can be mapped. Let $N_{syntax}(q_i)$ be the number of possible data model categories DM in which the query term q_i was observed to be mapped in a reference alignment corpus, and $count(q_i \rightarrow DM)$ the number of observed instances of the mapping to a specific alignment $q_i \rightarrow DM$. The probability of a term q_i syntactic mapping is given by:

$$P_{syntax}(Q) = \prod_{i=0}^{n} \frac{count(q_i \rightarrow DM)}{N_{syntax}(q_i)}$$

where $q_i \rightarrow DM$ are specific mappings. $H_{syntax}(Q)$ is computed by applying P_{syntax} into Shannon's entropy formula [13].

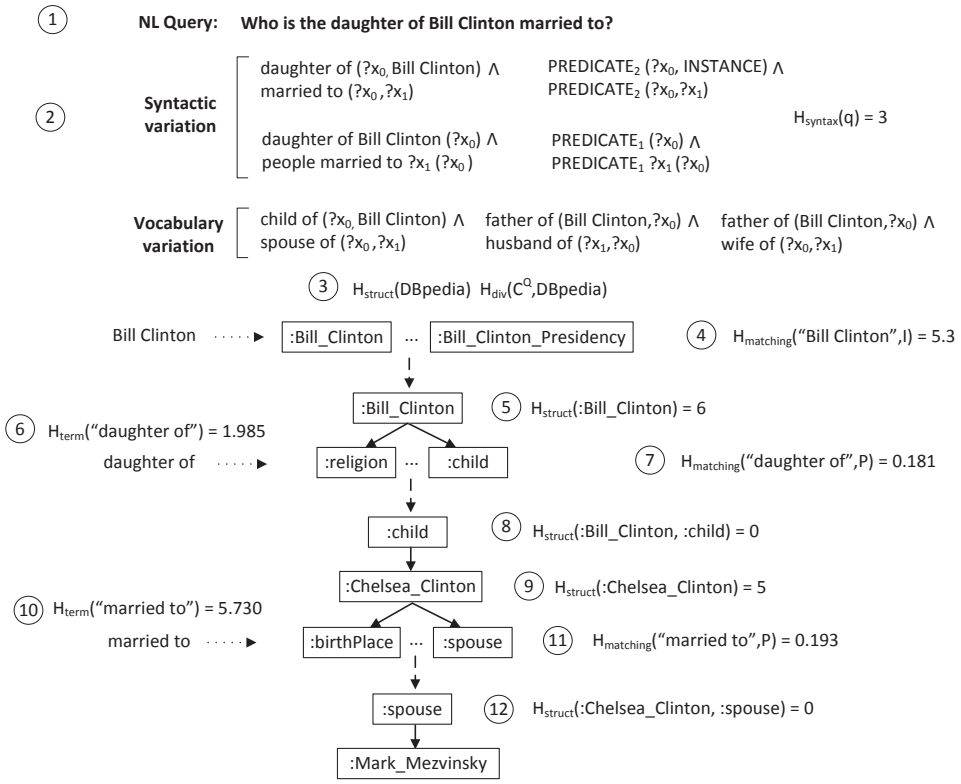

Figure 3: Instantiation of the query-entropy model for an example query.

4.1.2 Structural Entropy (H_{struct})

The *structural entropy* defines the complexity of a database based on the possible propositions that can be encoded under its schema. It provides a numerical description of the amount of information expressed in the database, independent of the query. Pollard & Biermann [11] proposed a structural entropy measure to quantify the entropy of a structured database. The entropy is computed by taking into account the number of predicates and constants and their syntactic combination. Figure 2 and Figure 3(5,8,12) depicts H_{struct}. The entropy of a constant c or a predicate π are defined as a function of the cardinality of the set of tuples in which the constant or the predicate is inserted. Further details are available in [11].

4.1.3 Terminological Entropy (H_{term})

The *terminological entropy* focuses on quantifying an estimate on the amount of *synonymy* and *vagueness* for the query or database terms. Let t be a query or database term containing the sequence of words $<w_0, w_1, ..., w_n>$. The terminological entropy is defined as a function of $P_{term}(w_i, w_j)$, i.e. the probability of a word w_i being expressed as a w_{j-th} related word for the associated ws sense (Figure 1). Under the query-database semantic matching problem, the relation between w_i and w_j can easily transcend the synonym relation, expressing a broader *semantic relatedness relationship*. Semantic relatedness will include both taxonomic (including different abstraction levels) and non-taxonomic relationships. As an absolute number of relationships cannot be enumerated, approximate entropy measures can be used to estimate the terminological entropy of a term. Figure 2 and Figure 3(6,10) depicts H_{term}.

One example of approximate terminological entropy measures is the *translational entropy* (Melamed, 1996) [10] which uses the coherence in the translation of a word (translational distribution) as an entropy measure. Given a set of word pairs of a set of ordered word pairs (s, t), respectively coming from a source language and a target language, an iterative process is used to determine the frequency $F(s, t)$ in which a word s is translated to a word t where $F(s)$ is the absolute frequency of the source word in the text. The probability that s translates to t is defined as $P(t|s) = F(s, t)/F(s)$. The notion of probability is defined by the translational distribution, the term $H(T|s)$ is generated, calculating the entropy of a

Measure	Semantic Measure Category	Type	Semantic Phenomena	Application
Pollard & Biermann [11]	Structural	Precise	Possibilities	Query-Data Alignment or Data
Translational Entropy (Melamed [10])	Terminological	Approximate	Ambiguity, Synonymy, Vagueness	Query or Data
Distributional Entropy	Terminological	Approximate	Ambiguity, Vagueness	Query or Data
Matching Entropy	Terminological	Approximate	Ambiguity, Vagueness	Query-Data Entity Alignments

Table 1: Classification of entropy measures according to associated features.

given word s against the target words set T: $H_{trans}(T|s) = -\sum_{t \in T} P(t|s)\ \log\ P(t|s)$.

4.1.4 Matching Entropy ($H_{matching}$)

Consists of measures which describe the uncertainty involved in the query-data matching/alignment between query terms and dataset entities. While terminological entropy measures provide an isolated estimate of the entropy, providing a prospective estimate of the matching complexity, the query-data matching entropy provides an estimate based on the set of potential alignments. These measures compute the uncertainty/ambiguity of an alignment under a semantic model Σ. Let q be an entity candidate in the query and let e_i be an i-th alignment candidate in the dataset. The *query-data matching entropy* can be estimated using the complement of a similarity metric $1 - sim_{space}(\overrightarrow{\mathbf{q}}, \overrightarrow{\mathbf{e_i}})$ such as cosine similarity, over a $word = \{w_0, \cdots, w_m\}$ or $concept = \{c_0, \cdots, c_n\}$ (e.g. distributional semantic model [8]) vector spaces. Distributional semantic models, semantic models based on the statistical patterns of co-occurrence of words within a large corpora can provide practical estimators for $H_{matching}$. Figure 2 and Figure 3(4,7,11) depicts the $H_{matching}$. In this case the entropy is not defined as a function of a probability but it is associated with a score.

5 Validation & Analysis

This section focuses on the validation and analysis of the proposed semantic complexity model. The model is validated using the Question Answering over Linked Data (QALD) 2011/2012 test collection [2], which is used as a challenge for the comparative evaluation of question answering systems over Linked Datasets. The performance of the participating Question Answering (QA) systems in addressing the schema-agnostic natural language queries is used as a gold standard for the validation of the semantic entropy model. The assumption is that queries with lower entropy positively correlate with the precision and recall performance of the system.

The QALD 2011/2012 test collections consist of 150 natural language queries over DBpedia 3.6 and DBpedia 3.7[1] as datasets. The QALD test collection was generated as a set of queries created by users around entities described in DBpedia. The set of questions covers different answer types and topics (e.g. proteins, countries, cities, companies, artists, planets, politicians, music, etc). QALD natural language queries explore different query patterns in the database.

The approximate entropy measures were setup using the following parameters:

- *Translational Entropy:* used the *European Parliament Parallel Corpus* for the generation of the translational corpus. The measure employed seven bitexts translating from English to Spanish, French, Portuguese, Italian, Greek, Swedish and Dutch, which were averaged to generate the final score.

[1] http://dbpedia.org/

- *Matching Entropy:* Generated as a set of vectors using the Explicit Semantic Analysis (ESA) [7] distributional semantic model over the Wikipedia 2013 corpus.

The *correlation* between each entropy measure and the *f-measure* of the participating QA systems was calculated taking into account the 150 queries in the test part of the QALD 2011 and 2012 test collections. Four top-performing QA systems were used in the evaluation: PowerAqua [9], Freya [4] for QALD 2011 and QAKis [1] and MHE for QALD 2012. The inter-annotator agreement between the PowerAqua [9] and Freya [4] is κ = 0.501 (95% confidence interval, 'moderate' agreement) and between QAKis [1] and MHE is κ = 0.236 (95% confidence interval, 'fair' agreement). A multiple linear regression model based on H_{syntax}, H_{term} (H_{trans}), $H_{matching}$ (H_{dist}) and H_{struct} was built.

The regression model parameters are shown in Table 2. H_{syntax} has a *significant negative correlation* with f-measure showing that the number of possible syntactical interpretations have a significant impact in the query interpretation process. Another *significant correlation is given by the terminological entropy measure* over the terms in the query which map to predicates (H_{term}, calculated by the translational entropy H_{trans}) The correlation shows that the translational entropy provide a valid estimator which reflect the higher level of ambiguity, synonymy and vagueness for predicate-type elements (the higher semantic gap for predicates is confirmed in Table 3). The $H_{matching}$ instantiated as H_{dist} *also presents a significant correlation for predicates*, confirming its suitability as an estimator for the vocabulary gap.

The *structural entropy H_{struct} of instances and classes showed a negative correlation with the f-measure*. The correlation is not significant for the structural entropy of the properties. This assymetry can be explained by the fact that in RDF the class or instance in most of the cases define the topic of the query (What is the highest *mountain*?, Who is the wife of *Barack Obama*?) having a *higher specificity* and being *more discriminative* in the definition of the data search space, while the properties tend to be more generic and reused across different contexts. The average structural entropy of instances (5.93) is significantly lower than the average structural entropy of properties (27.18). A query over a structurally more complex / better described entity (Barack Obama, with 505 associated triples) tend to be more difficult to resolve when compared to a less structurally complex entity (Michelle Obama, 268 associated triples).

Entropy Measure	**Estimate**	**Std. Error**	**t-value**	**Pr(>—t—)**
H_{syntax}	-0.05632	0.01697	**-3.317**	0.0011
H_{struct} Inst/Class (Sum)	0.00016	0.00599	0.027	0.97868
H_{struct} Prop (Sum)	-0.00013	0.00155	-0.086	0.93146
H_{trans} Pred (Sum)	-0.01330	0.01666	**-0.798**	0.42610
H_{dist} Pred (Sum)	-0.00202	0.00810	**-0.249**	0.80348

Table 2: Linear regression model between the evaluated entropy measures and the average f-measure of QA systems. Multiple R-squared = 0.1094 and adjusted R-squared = 0.0771

In addition to the entropy analysis, the queries were analyzed and categorized according to three dimensions (Figure 4): (i) query-term entity alignments, (ii) query features and (iii) query structure. This categorization supports a more in depth analysis of the impact of semantic complexity in the querying process.

All the 150 query-database alignments were analysed according to the type of their lexical alignment (*semantically related*, *similar string* (Dice coefficient >0.5), *substring*, *identical*). The distribution of query-database alignments is shown in Table 3. The proportion of *instances* which are *identical* to the query term is significantly larger compared to other categories, showing that the *lexical variability for instances (constants) is much smaller*. This is explained by the fact that instances usually map to named entities, which are less bound to synonymy, abstraction-level variations and vagueness. In contrast, *properties and classes (predicates)* tend to map to less specific terms, and *are more bound to ambiguity, synonymy and vagueness*. This is confirmed by the larger proportion of alignments for properties and classes under the *semantically related* category.

The queries were also analysed and categorized according to a set of query features: *contains in-*

Vocabulary Alignment Type	Vocabulary Type	Value
Semantically Related	Class	0.294
String Similar	Class	0.117
Identical	Class	0.117
Substring	Class	0.470
Identical	Complex Class	0.5
String Similar	Complex Class	0.1
Semantically Related	Complex Class	0.4
Semantically Related	Instance	0.098
Identical	Instance	0.696
Substring	Instance	0.147
String Similar	Instance	0.049
Missing Vocabulary Match	Instance	0.009
Missing Vocabulary Match	Null	1
Substring	Predicate	0.168
Missing Vocabulary Match	Predicate	0.109
Semantically Related	Predicate	0.411
Identical	Predicate	0.168
String Similar	Predicate	0.142
Identical	Value	0.25
Substring	Value	0.75

Table 3: Distribution of vocabulary gap types for each entity type (QALD 2011/2012).

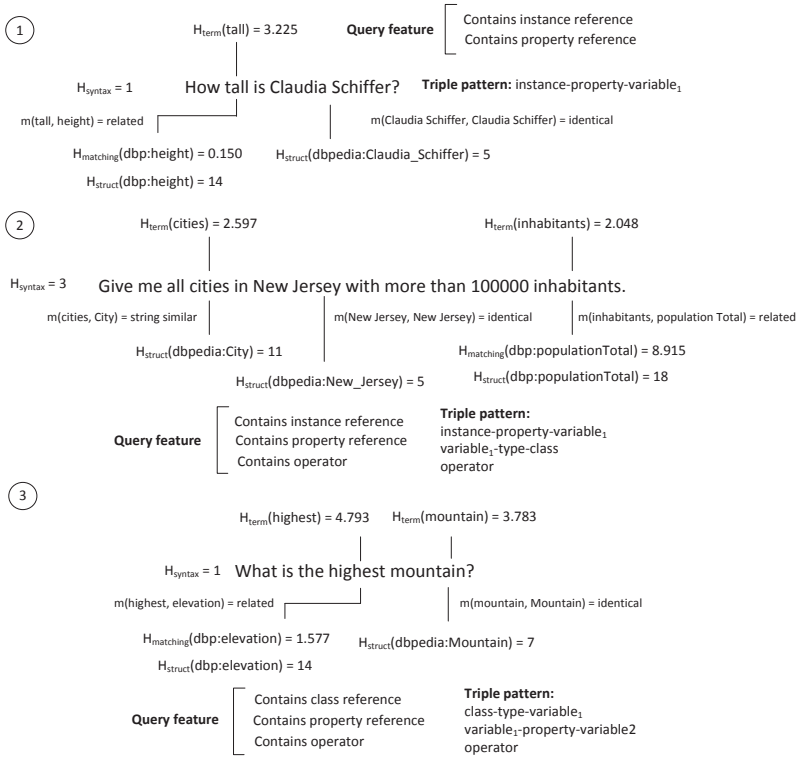

Figure 4: Entropy and query features for example queries.

stance reference, *contains class reference*, *contains property reference*, *contains complex class reference* (a complex class is a class with more than two words), *contains value reference contains operator reference*, *is a Yes/No question*. The features express the core natural language query - database mappings that need to be addressed by the query mechanism. The correlation between the query features and the average f-measure for the QA systems was also calculated and a multiple linear regression model was built (Table 4). Queries containing references to instances were positively correlated to the f-measure, while queries not containing instances were negatively correlated. This can be interpreted by combining this analysis with the alignment information from Table 3: predicate-type alignments are more bound to vocabulary variation (higher H_{term}, H_{vocab}) and are more difficult to resolve when compared to instance/value alignments.

Entropy Measure	Estimate	Std. Error	t-value	Pr(>—t—)
Instance	0.0750	0.1177	**0.638**	0.5247
Class	-0.0083	0.0816	-0.102	0.9189
Complex Class	-0.2118	0.1010	**-2.097**	0.0378
Property	0.0737	0.1659	0.444	0.6576
Value	-0.0565	0.1184	-0.478	0.6335
Yes/No	0.0054	0.1138	0.048	0.9615
Operator	-0.1506	0.0740	**-2.036**	0.0437

Table 4: Linear regression model between the query features and the average f-measure of QA systems. Multiple R-squared = 0.1171, adjusted R-squared = 0.09817.

Both *entropy* (H_{syntax}, H_{struct}, H_{term}, $H_{matching}$) and *query features* (*instances*, *complex classes*, *operators*) can be used as estimators for semantic complexity. Queries which were not or were poorly answered by the reference systems showed clear patterns which are correlated with entropy values: (i) high syntactic complexity (high H_{syntax}); (ii) high vocabulary gap (high $H_{matching}$, H_{term}) and (iii) predicate-based query (no instance reference in the query) (H_{struct}, H_{term}). Table 5 provides the classification of the set of unanswered/poorly answered queries according to the presence of high entropy values and also lists the non-trivial query term - database entity alignments. All unanswered queries fall into one (62%) or more (38%) of these categories.

5.1 Reflections on the Design of Schema-agnostic Query Mechanisms

The entropy measures and query feature analysis of the previous section can be used to define heuristics for maximizing the probability of a correct query-data matching in a schema-agnostic query scenario. A list of heuristics for addressing schema-agnostic queries are summarized below, based on the previous analysis:

1. **Prioritize the alignment of constants (instances):** Instances are less bound to vocabulary variation (lower H_{term}, H_{vocab}). The lower structural entropy H_{struct} associated with constants also allows the reduction of the search space.

2. H_{term} **can be used as a heuristic for matching complexity:** Having an estimation of the potential vocabulary variation of query terms predicates can be used to allow the prioritization of alignments with less ambiguity, synonymy and vagueness. H_{term} can be used to prioritize easier mappings.

3. H_{syntax} **is a strong estimator of query complexity:** Queries with complex compositional predicate patterns generate large entropy values which propagates to the matching stage. Schema-agnostic query mechanisms can explore query constraining approaches to minimize high H_{syntax} entropy values.

4. $H_{matching}$ **can be used as an estimator for the quality of the predicate alignment:** This value can be used to estimate the uncertainty of the alignment, supporting, for example, disambiguation mechanisms & clarification dialogs.

6 Conclusions & Future Work

This paper provides an analysis of measures of semantic complexity for schema-agnostic queries. A semantic model was built to understand the semantic dynamics behind the query-database semantic matching. Information theoretic models were used as a quantification model to measure the semantic complexity of mapping queries to database elements. The entropy measures and other query features were evaluated using a set of 150 natural language schema-agnostic queries over DBpedia by comparing the correlation between different Question Answering systems and the entropy measures. Syntactical,

Query	**Syntactic compl.** (H_{syntax})	**Vocab. gap** ($H_{matching}$, H_{term})	**Pred. Pivot** (H_{struct}, H_{term})	**Non-trivial alignments**
How many monarchical countries are there in Europe?		✓		monarchical countries - governmentType
Give me the capitals of all U.S. states.			✓	
Which states border Utah?		✓		border - east — border - southeast — border - south — border - northeast — border - north — border - west
Which mountain is the highest after the Annapurna?	✓	✓		highest - elevation
Which bridges are of the same type as the Manhattan Bridge?	✓	✓		type - design — type - design
Which state of the United States of America has the highest density?	✓	✓		highest density - densityrank
When did Germany join the EU?		✓		join - accessioneudate
Give me all soccer clubs in Spain.		✓		null - ground
Which German cities have more than 250000 inhabitants?	✓	✓	✓	inhabitants - populationTotal
How many students does the Free University in Amsterdam have?	✓			
What is the longest river?		✓	✓	longest - length
Does the new Battlestar Galactica series have more episodes than the old one?	✓			
Give me all people that were born in Vienna and died in Berlin.	✓	✓		died - deathPlace — born - birthPlace
Do Harry and William, Princes of Wales, have the same mother?	✓			
Give me all Australian nonprofit organizations.			✓	null - null
List all boardgames by GMT.		✓		null - publisher
Which countries are connected by the Rhine?				
Was the Cuban Missile Crisis earlier than the Bay of Pigs Invasion?	✓	✓		earlier - date
Give me all Frisian islands that belong to the Netherlands.	✓	✓		null - country
Which Greek goddesses dwelt on Mount Olympus?		✓		dwelt - abode
Which daughters of British earls died in the same place they were born in?	✓	✓		born - birthPlace — died - deathPlace
Who was called Scarface?		✓		called - nickname
Give me a list of all American inventions.		✓	✓	null - null
Which films starring Clint Eastwood did he direct himself?	✓			
Show me all songs from Bruce Springsteen released between 1980 and 1990.	✓	✓		songs - artist — release - releaseDate
Which movies did Sam Raimi direct after Army of Darkness?	✓			
What is the founding year of the brewery that produces Pilsner Urquell?	✓			founding year - foundation — brewery - brewery
Which country does the creator of Miffy come from?		✓		creator - creator — country - nationality
For which label did Elvis record his first album?		✓		null - releaseDate — label - recordLabel — null - artist —
% of unanswered questions	**51.7%**	**68.9%**	**20.6%**	

Table 5: 'Hard queries', i.e. queries which were nor or were poorly answered by the benchmarking systems. 'Checked' dimensions represent high entropy values.

terminological and matching entropies had a significant correlation with the results (f-measure) of the benchmarked systems. Based on the results, recommendations for the design of schema-agnostic query approaches were suggested. Future work will concentrate on the refinement of the entropy measures.

References

[1] Cabrio, E., Cojan, J., Aprosio, A.P., Magnini, B., Lavelli, A., Gandon, F.: Qakis: an open domain qa system based on relational patterns. In: Proceedings of the ISWC 2012. CEUR Workshop Proceedings, vol. 914 (2012)

[2] Cimiano, P., Lopez, V., Unger, C., Cabrio, E., Ngonga Ngomo, A.C., Walter, S.: Multilingual question answering over linked data (qald-3): Lab overview. In: CLEF (2013)

[3] Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 299–306. SIGIR '02, ACM, New York, NY, USA (2002)

[4] Damljanovic, D., Agatonovic, M., Cunningham, H.: Freya: An interactive way of querying linked data using natural language. In: Proc. of the European Semantic Web Conference Workshops. vol. 7117, pp. 125–138 (2012)

[5] Freitas, A.: Schema-agnositc queries over large-schema databases: a distributional semantics approach. In: PhD Thesis (2015)

[6] Freitas, A., Pereira Da Silva, J.C., Curry, E.: On the semantic mapping of schema-agnostic queries: A preliminary study. In: Workshop of the Natural Language Interfaces for the Web of Data (NLIWoD), 13th International Semantic Web Conference (ISWC) (2014)

[7] Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artifical Intelligence. pp. 1606–1611. IJCAI'07, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2007)

[8] Harris, Z.: Distributional structure. In: Word, 10(23). pp. 146–162 (November 1954)

[9] Lopez, V., Fernández, M., Motta, E., Stieler, N.: Poweraqua: Supporting users in querying and exploring the semantic web. Semantic Web 3(3), 249–265 (2012)

[10] Melamed, I.D.: Measuring semantic entropy. In: Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics. pp. 41–46 (1997)

[11] Pollard, S., Biermann, A.W.: A measure of semantic complexity for natural language systems. In: Proc. of the 2000 NAACL-ANLP Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems. pp. 42–46. Association for Computational Linguistics, Stroudsburg, PA, USA (2000)

[12] Popescu, A.M., Etzioni, O., Kautz, H.: Towards a theory of natural language interfaces to databases. In: Proceedings of the 8th International Conference on Intelligent User Interfaces. pp. 149–157. IUI '03, ACM, New York, NY, USA (2003)

[13] Shannon, C.: A mathematical theory of communication. Bell System Technical Journal 27, 379–423, 623–656 (1948)

[14] Sullivan, T.: Locating question difficulty through explorations in question space. In: Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries. pp. 251–252. JCDL '01, ACM, New York, NY, USA (2001)

[15] Tonellotto, N., Macdonald, C., Ounis, I.: Efficient and effective retrieval using selective pruning. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. pp. 63–72. WSDM '13, ACM, New York, NY, USA (2013)