Using Conceptual Spaces to Model Domain Knowledge in Data-to-Text Systems

Hadi Banaee and Amy Loutfi
Center for Applied Autonomous Sensor Systems
Örebro University
Örebro, Sweden
{hadi.banaee, amy.loutfi}@oru.se

Abstract

This position paper introduces the utility of the conceptual spaces theory to conceptualise the acquired knowledge in data-to-text systems. A use case of the proposed method is presented for text generation systems dealing with sensor data. Modelling information in a conceptual space exploits a spatial representation of domain knowledge in order to perceive unexpected observations. This ongoing work aims to apply conceptual spaces in NLG for grounding numeric information into the symbolic representation and confronting the important step of acquiring adequate knowledge in data-to-text systems.

1 Introduction

Knowledge acquisition (KA) is important for building natural language generation (NLG) systems. Two KA techniques including corpus-based KA and structured expert-oriented KA have been previously studied for NLG systems in (Reiter et al., 2003) to improve the quality of acquired knowledge. Both techniques use rule-based approaches in order to enrich the similarities between generated texts and natural human-written texts. An important class of NLG frameworks which use a rule-based approach is data-to-text systems where a linguistic summarisation of numeric data is produced. The main architecture of data-to-text systems has been introduced by Reiter (2007) which includes the following stages: signal analysis, data interpretation, document planning, microplanning and realisation. Domain knowledge for these systems is formalised as a taxonomy or an ontology of information. In a data-to-text architecture, all the stages are using the provided taxonomy. In particular, the signal analysis stage extracts the information that is determined in taxonomies such as simple patterns, events, and trends. Also, the data interpretation stage abstracts information into the symbolic messages using the defined taxonomies.

Most recent data-to-text frameworks have been developed using Reiter's architecture with the addition of providing the taxonomies or ontologies corresponding to the domain knowledge. For instance, the work on summarising the gas turbine time series (Yu et al., 2007) has used expert knowledge to provide a taxonomy of the primitive patterns (i.e. spikes, steps, oscillations). Similarly, the systems related to the *Babytalk* project (Portet et al., 2009; Gatt et al., 2009; Hunter et al., 2012) have stored medically known observation (e.g. bradycardia) in local ontologies. In order to avoid generating ambiguous messages, these systems simplify the stored information in the taxonomies by using only the primitive changes interesting for the end users. The core of such systems is still based on this fact - that the content of the generated text is dependent on the richness of the domain knowledge in the provided taxonomies which are usually bounded by expert rules. This organised domain knowledge is usually an inflexible input to the framework which restricts the output of the stages in data-to-text architecture. For instance, the taxonomy in (Yu et al., 2007) does not allow the system to represent unexpected observations (e.g. wave or burst) out of the predefined domain knowledge. Likewise, in the medical domain, an unknown physiological pattern will be ignored if it does not have a corresponding entity in the provided ontology by expert. This limitation in data-to-text systems reveals the necessity of reorganising domain knowledge in order to span unseen information across the data.

This position paper introduces a new approach, inspired by the conceptual spaces theory, to model information into a set of concepts that can be used by data-to-text systems. The conceptual spaces

Proceedings of the 8th International Natural Language Generation Conference, pages 11–15,
Philadelphia, Pennsylvania, 19-21 June 2014. ©2014 Association for Computational Linguistics

theory creates a spatial model of concepts that represents knowledge or information. This theory presents a promising alternative to modelling the domain knowledge in taxonomies or ontologies, particularly when a data-driven analysis is to be captured in natural language. This paper outlines the notion of conceptual spaces and illustrates how it can be used in a use case. Section 2 reviews the theory of conceptual spaces and its notions. Section 3 presents the approach for applying the conceptual spaces in NLG frameworks. In Section 4, a simple application of the proposed method is shown. Finally, we address the challenges and outline our plans for future work.

2 On the Theory of Conceptual Spaces

The idea of conceptual spaces has been developed by Gärdenfors (2000) as a framework to represent knowledge at the conceptual level. A *conceptual space* is formed in geometrical or topological structures as a set of *quality dimensions* describing the attributes of information to be represented. For instance, a conceptual space might comprise dimensions such as width, weight, or saltiness. A *domain* is represented to be a set of interdependent dimensions which cannot logically be separated in a perceptual space. A typical example of a domain is ‘colour’ which can be defined through multi dimensions like hue, saturation, and brightness. *Properties* are the convex regions in a single domain describing the particular attributes of the domain. As an example, ‘green’ is a property corresponding to a region in the colour domain (Fig. 1, right). In natural language, properties are mostly associated with adjectives in a particular domain. A conceptual space contains a membership distance measure for each property within the domains which represents the regions occupied by the property and allows to depict the notion of similarity (Rickard et al., 2007).

Concepts are formed as regions in a conceptual space. In particular, a concept is represented as a set of related properties which might cover multiple domains together with information how these domains are correlated. For instance, the concept of ‘apple’ can be represented as regions in colour, size and taste domains (Fig. 1). The representation of concepts in space contains an assignment of weights to the domains or dimensions, in order to distinguish between similar concepts (Gärdenfors, 2004). In natural languages, concepts often correspond to nouns or describe verbs when time is involved as a dimension (Rickard et al., 2007). The most representative instance of a concept is its *prototypical member* which is represented as an n-dimensional point in the concepts region. The conceptual space can be geometrically divided (e.g. using *Voronoi tessellation* (Gärdenfors, 2004)) to a set of categories corresponding to the prototypical members. *Objects* (such as instances, entities, or observations) in a conceptual space are identified in the concept regions which characterised as vectors of quality values. For example, a particular instance of ‘apple’ is depicted in Fig. 1 as a vector of properties <green, medium, sweet–sour>. An object contains a property depending on the nearness of its point to the defined region of the property. This notion leads to have a similarity measure within a domain to identify the properties of objects. *Similarity* is an essential notion in any conceptual space framework which is defined on individual domains. The geometrical representation of conceptual spaces provides the ability of using distance measures, which is missed in purely symbolic representations, to consider the similarity of concepts and instances.

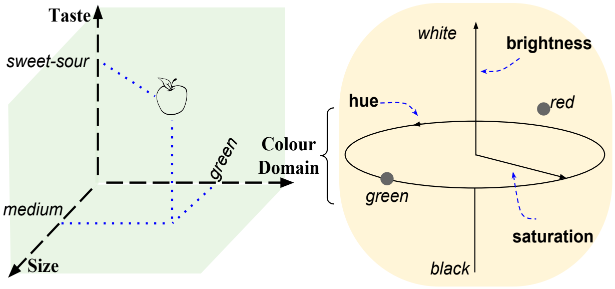


Figure 1: A typical example of a conceptual space to represent ‘apple’ concept.

3 Proposed Approach: Conceptual Spaces for Data-to-Text Systems

This section describes the usage of conceptual spaces for modelling numeric knowledge as concepts into a spatial representation. The proposed approach shows how to use conceptual space theory to reorganise the predefined taxonomies into a set of concepts in order to represent unexpected patterns. The idea consists of two phases, constructing a conceptual space corresponding to the taxonomy, and enhancing the regions in the space based on new observations. The general steps of the proposed approach are described as follows:

Step 1: Build the required taxonomy of observations and patterns in the same way as traditional data-to-text systems in order to provide a set of primitive information requirements using the expert-oriented, domain, or corpus-based knowledge. Primitive entities from these taxonomy will be the n-dimensional vectors of concepts in conceptual space.

Step 2: Initialise a conceptual space and determine its components, including quality dimensions, domains, and concepts corresponding to the domain knowledge and the context of data. Using similarity measures on the determined dimensions, the model is able to define the geometrical distance between each pair of vectors and identify the nearest concept for any point in space. By defining the applicable domains and dimensions, the conceptual space is able to characterise a vast range of interesting concepts, which may not be similar to the provided entities.

Step 3: Specify the ontological instances gathered in step one as concepts regions. This step grounds the primitive observations to a set of prototypical members as n-dimensional vectors in the created conceptual space. Also the space would be classified into a set of categories presenting the properties of the prototypical members. The main contribution of this approach is based on the fact - that by providing the semantic information as geometrical vectors, the model is spanned to conceptualise the information categories which enables calculating the similarities between knowledge entities like new (non-primitive) extracted patterns as new vectors in the space. However, a new entity could be 1) close to an existing prototypical member and placed in its geometrical category, or 2) an anomalous point and placed as a new prototype in the space.

Step 4: Rearrange the conceptual categories corresponding to the prototypical members by adding new instances to the model as new vector points. The symbolic properties of prototypical members in space are used to describe novel properties of unknown entities. When a new observation appears in space as a vector, it leads to reorganise the boundaries of concepts regions related to the new inserted member. The expanded space will provide more descriptive regions for unconsidered entities. It is notable that the provided domains and dimensions enables the conceptual space to grow with new entities which are event sans association with existing categories.

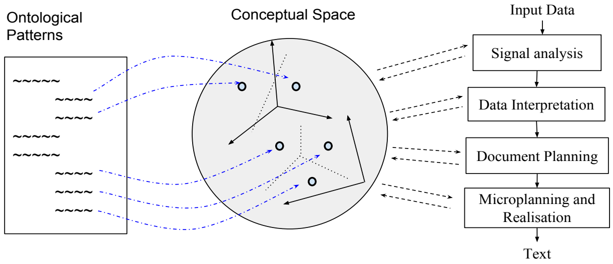


Figure 2: The conceptual space in data-to-text architecture as an alternative for ontological patterns.

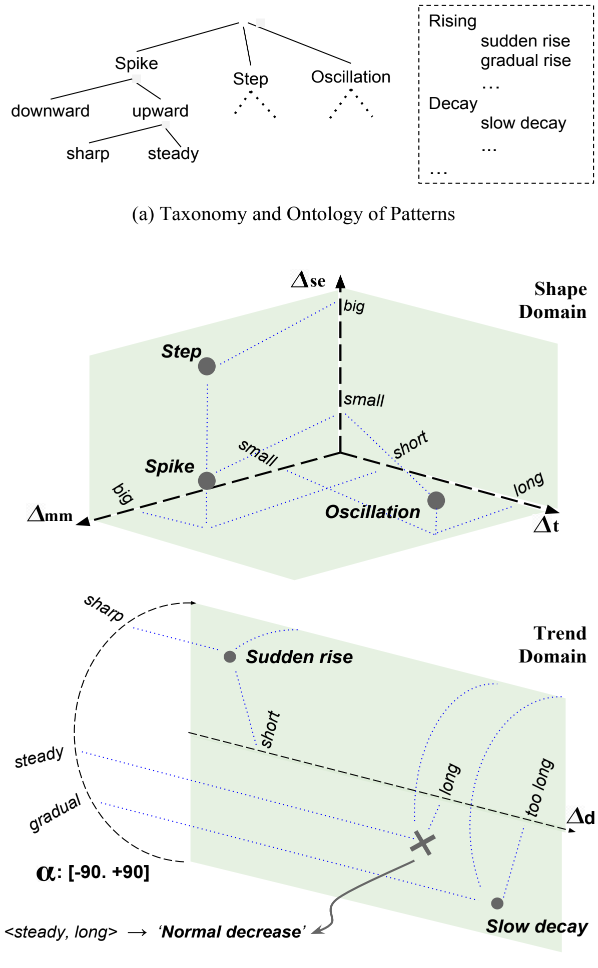
Different stages of data-to-text architecture can be connected to the built conceptual space instead of their relations to the ontology. Specifically, pattern discovery in the signal analysis stage does not need to be limited to rules and domain constraints.

Data-to-text approaches which use ontologies for signal processing are able to apply probabilistic or fuzzy processes to map the patterns of data into the "*most likely*" concepts in ontology. However, one advantage of the proposed approach is that enables the system to represent new concepts that are non-relatively deviant cases, as well as covering intermediate patterns. So, any extracted information from data can be formalised in the conceptual space and then be characterised in a symbolic representation. Another advantage of this model is that the conceptual space assists the system to enrich the quality of represented messages in the final text with considering unseen, but interesting information for the end users. Fig. 2 depicts the conceptual space in relation with the stages of the data-to-text architecture.

4 Use Case: From Data Streams to Conceptual Representation

Knowledge extraction in data streams exploits the most informative observations (e.g. patterns and events) through the data (Rajaraman et al., 2011). In most of data-to-text systems, much attention has been given to the sensor data as the best indicator of data streams (e.g. weather sensor channels, gas turbine time series, and physiological data in body area networks). A robust text generation system for sensor data needs to provide a comprehensive information structure in order to summarise numeric measurements. Here, we explain how the proposed approach can apply to model the defined taxonomies in sensor data applications, particularly for gas turbine time series (Yu et al., 2007) and neonatal intensive care data (Gatt et al., 2009). The main challenge here is the definition of concepts and quality dimensions from non-sensible observations in time series data. However, a preliminary model is introduced as follows:

Based on the acquired knowledge in both systems, the patterns are categorised to 1) primitive disturbance shapes: spikes, steps, and oscillations, or 2) partial trends: rise, fall, and varying. These observations are associated with a set of attributes and descriptions for their magnitude, direction and/or speed (e.g. downward, upward, or rapidly, normally, etc.). A typical demonstration of taxonomies/ontologies in traditional data-to-text systems dealing with sensor data has been shown in Fig. 3-a. Our method exploits these structures to build an applicable conceptual space related to the acquired knowledge. It is worth noting that building the components of the conceptual spaces for different sensor data in other contexts would differ. To cover the observations in time series, two domains are defined: shape and trend domains. For the shape domain, the rules behind the definition of primitive events lead to determine quality dimensions. For instance, ‘spike’ is defined as “small time interval with almost same start and end, but big difference between max and min values”. So, the spike concept can be characterised in the shape domain by quality dimensions: *time interval* (Δt), *start-end range* (Δse), and *min-max range* (Δmm). The prototypical member of spike concept can be represented as a vector of properties: v_1:<short Δt, small Δse, big Δmm>. Same dimensions can describe the steps and oscillations, shown in Fig. 3-b (top). For the trend domain, finding descriptive dimensions and properties is dependent on the selected features in the trend detection process (Banaee et al., 2013). Here, the provided quality dimensions for the trend domain include: *trend orientation* (α), and *trend duration* (Δd). As an example, ‘sudden rise’ concept can be represented as a region in the trend domain with a prototypical member vector v_2:<positive sharp α, short Δd>, shown in Fig. 3-b (bottom). The complex concepts can be spanned to multi domains with their properties regions. For instance, ‘rapid upward spike’ pattern is definable as a region in space, spanned in both shape and trend domains, which its representative vector has five property values in all dimensions like: v_3:<v_1, v_2>.



(a) Taxonomy and Ontology of Patterns

(b) Conceptual Space (Shape and Trend domains)

Figure 3: A conceptual space proposed for modelling domain knowledge in sensor data. a) Taxonomy and ontology of patterns, b) Shape domain and trend domain.

This modelling has an effect on signal analysing in that any unseen event and trend can be extracted and represented by finding the nearest prototypical instances in the corresponding vector space. Fig. 3-b (bottom) depicts an example of two points represented ‘sudden rise’ and ‘slow decay’ trends in the space. The location of a new instance in space, e.g. <steady, long> is computable by calculating geometrical distances of their properties, and consequently the corresponding descriptive symbols can be inferred as ‘normal decrease’.

This use case focuses on event-based observations based on the shapes and trends of patterns in sensor data. Other contexts may be interested to represent other observations like repetitive rules, motifs and unexpected trends which need particular studies on how to model these issues in conceptual spaces and capture their properties.

5 Discussion and Conclusion

This position paper has presented the notion of conceptual spaces as an alternative approach to modelling domain knowledge in data-to-text systems. The next obvious steps are to use conceptual spaces in a NLG framework and experimentally validate their suitability for capturing data-driven events, patterns, etc. This paper has attempted to motivate the use of conceptual spaces in order to cope with information which cannot be accurately modelled by experts. Still, however, some remaining challenges are to be addressed. One challenge is determining a comprehensive set of domains and quality dimensions representing the acquired knowledge in a conceptual space. Another challenge is grounding concepts to linguistic description in order to provide a thorough symbolic description of quantitative vectors in the space. A further challenge is lexicalisation in modelling the conceptual spaces, which is related to choosing accurate words for the conceptual regions regarding to the semantic similarities for properties of the concepts, without using expert knowledge.

Acknowledgments

The authors of this work are partially supported by SAAPHO project: Secure Active Aging: Participation and Health for the Old (AAL-2010-3-035).

References

Ehud Reiter, Somayajulu G. Sripada, and Roma Robertson. 2003. Acquiring Correct Knowledge for Natural Language Generation. *Journal of Artificial Intelligence Research*, 18:491–516.

Ehud Reiter. 2007. An architecture for data-to-text systems. *ENLG'11: the Eleventh European Workshop on Natural Language Generation*, 97–104.

Jin Yu, Ehud Reiter, Jim Hunter, and Chris Mellish. 2007. Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, 13(1):25–49.

François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7):789–816.

Albert Gatt, François Portet, Ehud Reiter, Jim Hunter, Saad Mahamood, Wendy Moncur, and Somayajulu Sripada. 2009. From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management. *AI Communications*, 22(3):153–186.

James Hunter, Yvonne Freer, Albert Gatt, Ehud Reiter, Somayajulu Sripada, and Cindy Sykes. 2012. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. *Artificial Intelligence in Medicine*, 56(3):157–172.

Peter Gärdenfors. 2000. *Conceptual Spaces: The Geometry of Thought.* MIT Press.Cambridge, MA.

John T. Rickard, Janet Aisbett, and Greg Gibbon. 2007. Reformulation of the theory of conceptual spaces. *Information Sciences*, 177(21):4539–4565

Peter Gärdenfors. 2004. Conceptual spaces as a framework for knowledge representation. *Mind and Matter*, 2(2):9–27.

Anand Rajaraman, and Jeffrey D. Ullman 2011. *Mining of massive datasets.* Cambridge University Press.

H. Banaee, M. U. Ahmed, A. Loutfi 2013. A Framework for Automatic Text Generation of Trends in Physiological Time Series Data. *SMC'13: IEEE International Conference on Systems, Man, and Cybernetics*, 3876–3881.