Technical term similarity model for natural language based data retrieval in civil infrastructure projects

T. Le and H. D. Jeong

Department of Civil, Construction and Environmental Engineering, Iowa State University, United States
E-mail: ttle@iastate.edu, djeong@iastate.edu

Abstract –

Recent advances in data and information technologies have enabled extensive digital datasets to be available to decision makers during the life cycle of a civil infrastructure project. However, much of the data is not yet fully reused due to the challenging and time consuming process of extracting the desired data for a specific purpose. Digital datasets are presented only in computer-readable formats and they are mostly complicated. In order to accurately extract a required subset of data, end users need to have deep understanding of the structure of the data schema, the meaning of each data entity and a query language. Thus, to truly facilitate the reuse of digital project data, a computational platform is needed to allow users to present their data needs in natural language. One of the critical requirements for a computer to perform this task is the ability to understand and interpret users' natural language inputs where keywords are a basic linguistic component. This research aims to collect technical terms commonly used in the civil infrastructure domain and develop a semantic similarity model that can measure the meaning relatedness/similarity between terms. Natural Language Processing (NLP) techniques and C-value method are used to automatically extract terms from text documents. A machine learning model called Skip-gram model is then employed to learn the semantic relatedness between technical terms using an unlabeled highway corpora as the input data. The input corpus includes 10 million words mainly collected from roadway design guidelines across the U.S. The model is evaluated by comparing the mapping results performed by a computer and a human.

Keywords –

Civil infrastructure project, Landxml, Data retrieval, Natural language interface, NLP, Vector space model

1 Introduction

The advanced computerized technologies such as 3D modeling and Geographic Information System (GIS) throughout the life cycle of a civil infrastructure project has allowed a large portion of project data to become available in digital format. In order to enable digital data exchange between proprietary software applications, several neutral data standards, for instance, LandXML [1] and TransXML [2], have been developed. However, these schemas are presented in machine-readable format and so complicated that it is difficult for end users to extract the desired properties [3]. The end user is required to have considerable programming skills and properly understand the structure and the meaning of each entity or attribute included in the source data schema. Thus, there have been apparent demands for an automatic data extraction means that would eliminate manual processing.

To address the above demand, a considerable amount of research efforts has been undertaken in both the building and transportation sectors. The most commonly adopted method is Model View Definition (MVD) which defines a subset of data for a specific business process. Some examples from this line of efforts include Construction to Operation Building Information Exchange (Cobie) [4] for IFC schema and InfraModel subsets for LandXML schema. Although a large number of MVDs have been proposed, they have not yet kept up with the dynamic demand from the industry since the current method for MVD development is based on a manual process which is highly time consuming [5,6,7]. Moreover, the business processes are dynamic and tend to change over time, hence MVDs must be periodically maintained and tailored to reflect the changes from industry practices. Therefore, there is a need to change the current practice of model view definition from the ad-hoc approach to a more rigorous methodology [5].

A natural language interface that can allow for human-computer interaction in natural language would enable digital data retrieval to overcome the bottle-neck of MVDs and remove the current burden on the end user.

One fundamental requirement for such a system is the ability to understand technical terms/keywords since they are a basic unit of natural language and users prefer to use them for obtaining data [8]. One of the major obstacles to fulfill the above requirement is the ambiguity issue of technical terms. A technical term in a domain specific document implicitly refers to something that only experts in that field can correctly understand. For example, the term 'roadway type', in general context, can mean the classification of roadways in terms of either material, function or location; but in the highway context, it refers to roadway functional classification. Another issue related to term ambiguity is that two different terms may be used to represent the same concept. For instance, the concept of longitudinal centerline of a roadway has a variety of terms including 'profile', 'crest', 'grade-line' and 'vertical alignment'. Addressing those issues will provide a foundation for natural language interfaces to fast and exactly extract data from the complicated sets of data with minimized human intervention and costs.

To fulfill that need, this research aims to propose a novel model that can be used to measure the semantic similarity between technical terms in the civil infrastructure domain. In order to achieve that goal, Natural Language Processing (NLP) techniques and C-Value method [9] are employed to process domain-specific guidelines and extract technical terms commonly used in the civil sector. A matching algorithm implementing the result from the previous step is developed to automatically look for the nearest entities and attributes in the Landxml schema for a certain keyword. The proposed semantic similarity model and the data mapping algorithm are evaluated by comparing the automatically retrieved data with the results performed by a human for performance assessment.

2 Related research

2.1 Partial digital model extraction

Methods for extracting partial models for specific use cases can be classified into the following three groups ordered by the degree of ease of use for end users: (1) developing a query language specifically for Building Information Modeling (BIM) models, (2) ontology-based query approaches, and (3) user-oriented query methods. The first group aims to tailor the conventional query languages (e.g., SQL, Object Orientation) for extracting information from BIM models. The major focus is on developing spatial filter strategies. Examples of these efforts include the Spatial query language [3], QL4BIM Spatio-semantic query language [10]; graph-based BIM retrieval [11], and topological querying [12]. The second group is to enhance the human-readability of data schema by utilizing an ontology approach to transform relations among data entities from implicit to explicit. With these semantic representations, it is easier for end users to read and comprehend a complicated data schema. An extensive number of studies based on this approach have been carried out for various use cases including ontology-driven construction information retrieval for tunnel projects [13], ontology partial BIM model extraction for building projects [14], ontology-based extraction of construction information [15] and ontology based querying over linked life cycle data spaces [16]. The last class of partial model query approaches moves a step further in terms of enhancing the ease of data extraction by providing query tools that require less effort from users. For example, Won et al. (2013) [17] proposed a non-schema algorithm that allows for the extraction of IFC instances without using IFC schema or MVD. In addition, a visual BIM query [18] was also established to visualize query codes. Although significant research efforts have been conducted, there is still a lack of natural language interface platforms that can enable computers to understand and interpret the end user's data interests in the civil infrastructure domain.

2.1.1 Semantic data label matching

In the construction industry, research efforts are currently focusing on standardizing the data structure format. There are very few studies that dealt with the issue of sense ambiguity. Zhang and El-Gohary (2015) [19] proposed an algorithm called ZESeM aiming to match a certain keyword to the most semantically relevant IFC entity. The algorithm includes two sequential steps including term-based matching and semantic relation based matching. Since the algorithm accepts matches from the label-based matching step, disambiguation still remains where the same word form is used for different senses. In addition, since ZESeM relies on Wordnet which is a generic lexicon, the applicability would be limited. Lin et al. (2015) [20] developed an IFD based framework for BIM information retrieval. IFD Library (International Framework for Dictionaries library), which is developed and maintained by the international buildingSMART, is a dictionary of BIM data terminology that assigns the same ID to synonyms. The integration or exchange of data using IDs rather than data names would eliminate semantic mismatching. However, since IFD is a hand-made electronic vocabulary, constructing this e-dictionary is time consuming and therefore it is still very limited considering the large collection of terms used in the construction industry.

2.2 Natural Language Processing

NLP is a collection of techniques that can analyze and extract information from natural language like text and speech. The major applications of NLP include translation, information extraction, and opinion mining [21]. These applications are supported by a combination of several techniques such as Named Entity Recognition (NER), Part-of-Speech (POS) tagging [22,23], tokenization (or word segmentation) [24,25], relation extraction, sentence parsing, word sense disambiguation [26,27,28], etc. NLP methods can be classified into two main groups: (1) rule-based and (2) machine-learning (ML) based methods. Since the former group is based solely on hand-coded rules, these systems are not able to cover all the complicated set of human grammars [29]; and therefore they do not perform well. NLP research has shifted from rule-based analysis to ML-based methods [21]. ML models are able to learn patterns from training examples and predict the output, hence they are independent to languages, linguistic grammars and consequently human resource cost would be reduced [30].

2.3 Word vector representation

Vector space model is a popular method used for estimating word similarity when digital dictionaries are not available. This method is based on the distributional model which represents meanings of words through their contexts (surrounding words) in the corpus [31]. The distributional model stands on the distributional hypothesis that states that two similar terms would occur in the same context [32]. The outcome of this approach is a Vector Space Model (VSM) in which words are converted into vectors and the similarity between two vectors represents the context similarity between the corresponding words [31]. VSM outperforms the dictionary-based method in terms of time saving as the semantic model can be automatically obtained from text corpus and collecting of these corpus is much easier than manually constructing a digital dictionary [33]. Among the methods to develop VSM, Skip-Gram model [34], which is an un-supervised machine-learning model, outperforms other statistical computational methods in various performance aspects such as accuracy and degree of computational complexity [34]. This machine-learning model learns the semantic similarity between two technical terms through their context similarity. The outcome of the training process is a set of representation vectors for technical terms.

3 Highway term space model development

The ultimate goal of this research is to build a highway vector space model (H-VSM) that can support the disambiguation task for advancing the computer-human interaction in partial model extraction. For addressing ambiguity, there are several methods including thesaurus based, ontology based and distributional method. The first two methods require a full lexicon or ontology consisting of semantic descriptions for all relevant concepts. These methods would be ideal for the disambiguation task if domain related thesauruses are available. However, since building those dictionaries requires a huge amount of empirical work, they can cover only limited vocabulary. Wordnet is one of the largest lexicons available containing 117,000 synsets [35], but it is generic and not suitable for the highway domain. For this reason, this research employs an unsupervised machine learning method called Skip-gram to train unlabeled data and learn the meanings of words by analysing their context words. The process of developing the H-VSM includes the following steps: (1) text document collection, (2) technical term extraction, and (3) semantic similarity training. The sub-sections below discuss the detailed procedures for each step.

3.1.1 Data collection

In this study, a highway corpus has been built using technical documents including textbooks, and highway engineering manuals collected from the Federal Department of Transportation (DOT) and from 18 State DOTs. The focus of the highway corpora in this research is on three project phases including design, construction and asset management. Technical terms in a guidance document in the engineering field are organized in various formats such as plain text, tables, and equations. Since tables and equations are not yet supported by the state-of-the-art NLP techniques, they are removed from the text corpora. The result of data collection is a plain text corpora consisting of approximately 10 million words. This dataset is utilized to extract highway related technical terms which are then trained and converted into vectors.

3.1.2 Technical term detection

The first step of data processing to construct the H-VSM is to detect highway related technical terms. In order to achieve this goal, a set of NLP techniques including tokenization, part of speech tagging are utilized to identify the POS tag for each word in the highway corpora. The OpenNLP library is used to perform this task. The linguistic process, as illustrated in Figure 1, includes the following steps:

- ***Word Tokenization:*** In this step, text is broken down into individual units (called tokens).
- ***POS tagging:*** The purpose of this step is to determine the part of speech tag (e.g., noun, adjective, verb, etc.) for each token.
- ***Noun phrase detection:*** Linguists argue that a technical term is either a noun (e.g., road) or a noun phrase (NP) (e.g., right of way) that frequently occurs in the domain text documents [36]. This research utilizes the following patterns $(A|N)_1{*}N_1$ Prep(of) $(A|N)_2{*}N_2$ or $(A|N)_1{*}N_1(A|N)_2N_2$ to detect good candidates for technical terms. In the patterns above, A is adjective and N is noun.
- ***Termhood measurement and concept extraction:*** The C-Value algorithm, proposed by Frantzi et al. (2000) [9], is employed to rank the candidates extracted from the previous step. C-value represents the degree of termhood and is computed based on the frequency of occurrence and the length of terms. Equation 1 below presents the C-value measurement of termhood.

$$C\_value(a) = \begin{cases} log_2|a|f(a), & a\ is\ not\ nested \\ log_2|a|f(a) - \frac{1}{P(Ta)}\sum_{b\in Ta} f(b), & otherwise \end{cases} \quad (1)$$

Where: a is a candidate noun phrase, f is the frequency of a in the corpus, Ta is the set of extracted noun phrases that contains a, and P(Ta) is the number of these candidate terms.

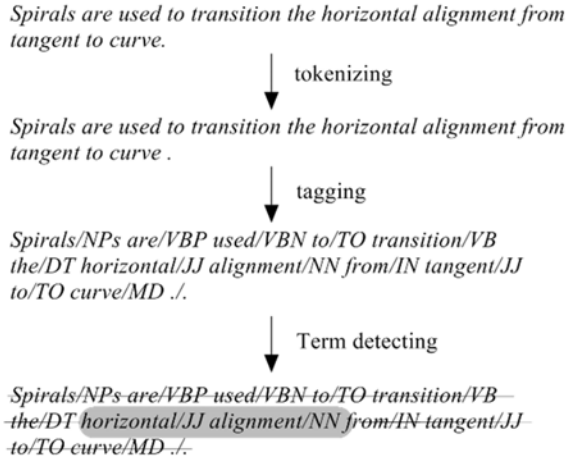

Figure 1. Term detection procedure

3.1.3 Data training and results

The tagged text corpora from the previous phases will serve as a data source for developing the semantic similarity model. Before collecting the training datasets, multi-word terms in the corpus are replaced with connected blocks of their word members so that they can be treated as single tokens. For instance, 'vertical alignment' will become 'vertical_alignment'. To train the highway corpus, this research employs the skip-gram neural network training model which was developed by [37]. The Skip-Gram model, as illustrated in Figure 2, requires a set of training data in which the input data is a linguistic unit (noun or noun phrase) and the output data is a set of context words. In order to collect this training dataset, the tagged text corpus is scanned to collect instances of terms and their corresponding context words. Each occurrence of a technical term will correspondingly generate a data point in the training dataset.

The semantic similarity is trained using the word2vec package developed based on the Skip-gram neural network model. Parameters required for word2vec include frequency threshold values, hidden layers, and context window. The parameters used for the training model are presented in Table 1.

Table 1. Skip-gram model parameters

Parameter	Value
Frequency threshold	50
Hidden layer size	300
Context window size	10

Figure 3 presents the term space model developed from the training process. In this model, each technical term collected from technical documents is represented as a vector in a high dimensional space; and the distance between them represents the semantic similarity. The preliminary term space presented in this paper consists of more than six thousand technical keywords.

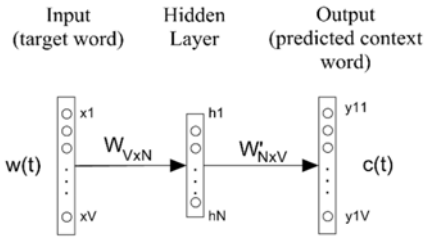

Figure 2. Skip-gram model

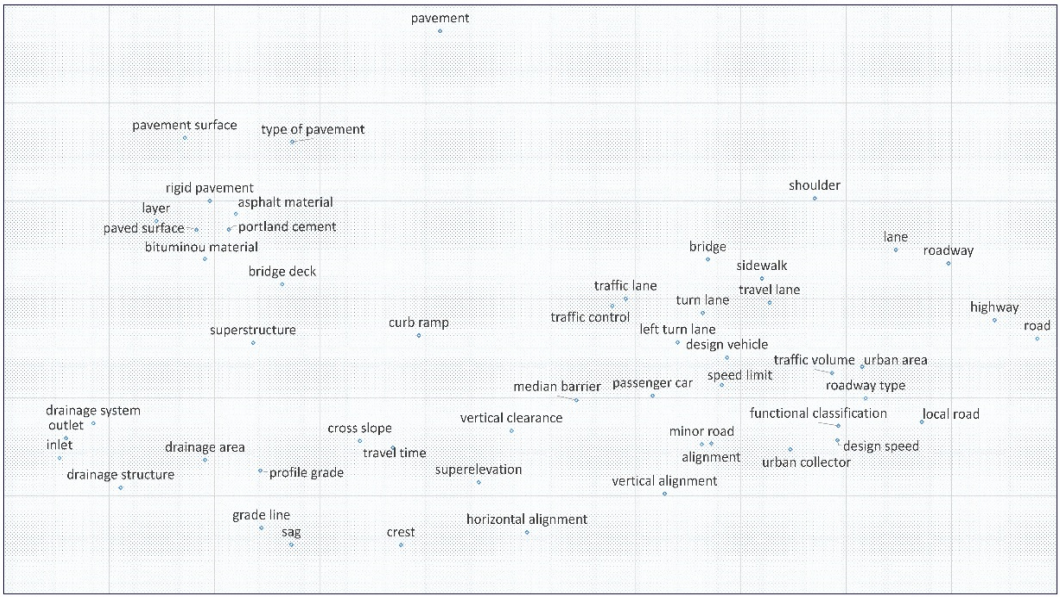

Figure 3. Highway term space model (H-VSM)

4 Test and evaluation

4.1.1 LandXML entity search

The proposed semantic model has been tested on the ability to search for data entities in the Landxml schema. As part of the test, an algorithm (see Figure 4) was developed for semantically searching for equivalent entities/attributes in Landxml schema. The algorithm is a two-stage procedure. In the first stage, a list of the nearest terms of the keyword input is generated by utilizing the vector space model developed in this research. A string-based similarity algorithm is then applied to find the entity in the Landxml schema that has the most similar name for each synonym in the list. The string similarity between a word in the nearest list and a Landxml entity or an attribute is computed using the Levenshtein algorithm [38] which is an edit-distance matching based method. The final matches are ranked by the similarity score.

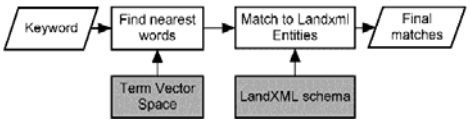

Figure 4. Landxml entities search algorithm

4.1.2 Evaluation

An evaluation experiment was conducted to evaluate the performance of H-VSM and the searching algorithm. In this experiment, a graduate student was asked to look for the data entities and attributes in the Landxml schema that are equivalent or the most related to each of 37 randomly selected keywords. Meanwhile, a prototype built upon the developed algorithm was applied to automatically generate the most matched Landxml entities. The results from the two methods were used to calculate the accuracy of the searching algorithm.

$$Recall = \frac{correctly\ matched\ keywords}{total\ keywords} \quad (2)$$

$$Precision = \frac{correctly\ matched\ keywords}{total\ matched\ keywords} \quad (3)$$

$$F_{measure} = \frac{2.Precision.Recall}{Precision + Recall} \quad (4)$$

Table 2 presents a portion of the experiment result and Table 3 shows the evaluation result using the criteria presented in Equations 2, 3 and 4. As presented in Table 3, the system shows a 37 percent precision which is relatively low. This is possibly due to the training data size. The searching algorithm accuracy highly relies on the capacity of finding synonyms which is based on the vector space model. This model is currently based on a data training set consisting of only approximately 10 million words. In order to enhance the accuracy, the data training set needs to be extended. Future research will be conducted to extend the training data set.

Table 2. A portion of the experiment result

Keyword	Top matched Landxml entities	Y/N
Drainage system	Outlet OutletStruct	Yes
Vertical alignment	roadTerrain pointGeometry	No
Pavement type	pavementSurfaceType stateType	Yes
Roadway type	Project type Classification Roadsign type	Yes

Table 3. Evaluation result

Recall (%)	Precision (%)	F-measure (%)
27	37	31

5 Conclusions

Digital project data is now widely available throughout the project life cycle in the civil infrastructure sector. However, the data collected and generated in the early project development stages are not typically reusable in the downstream phases. This is due to the interoperability issue when digital data from the original data creator is not readable or correctly understandable by the data receiver. This research developed a framework that semantically searches for the desired data from a transferred data file. The framework is composed of two components including (1) a term space model which represents highway related concepts extracted from the highway corpora in vectors and (2) a searching algorithm that can search for entities in the Landxml schema based on their semantic similarity instead of string based similarity.

The framework was evaluated by testing on a randomly selected set of input keywords. The result shows the accuracy of over 30 percent. The accuracy is low due to the size of the training data. Future research will be conducted to increase the data size.

The developed semantic similarity model is expected to serve as a fundamental resource for data integration and query systems to eliminate the issue of data terminology inconsistency. The model also has a broad impact in text mining for the infrastructure sector. Although digital data is increasingly available, text documents are still a major data and information communication channel among project stakeholders. The ability to understand the meanings of technical terms will enable computer systems to fast and exactly extract value information from project documents such as contracts or reports. Thanks to that, laborious work of reading and manipulating data and information in paper format would be eliminated.

References

[1] landxml.org, About landxml.org, On-line: http://www.landxml.org/, Accessed: 10/11/2005.

[2] Scarponcini P., Methodology for selection and development of transxml schemas, *Transportation Research Record: Journal of the Transportation Research Board* (1) 107-115, 2008.

[3] Borrmann A., Rank E., *Query support for bims using semantic and spatial conditions*, Handbook of Research on Building Information Modeling and Construction Informatics: Concepts and Technologies, 2009.

[4] East E.W., Construction operations building information exchange (cobie), *Tech. rep., DTIC Document*, 2007.

[5] Venugopal M., Eastman C. M., R. Sacks, J. Teizer, Semantics of model views for information exchanges using the industry foundation class schema, *Advanced Engineering Informatics* 26 (2) 411-428, 2012.

[6] Eastman C., The future of ifc: Rationale and design of a sem ifc layer, *presentation at the IDDS workshop*, 2012.

[7] Hu H., Development of interoperable data protocol for integrated bridge project delivery, Ph.d., *UMI Dissertations Publishing 2014*, 2014.

[8] Shekarpour S., Auer S., Ngomo A.-C. N., Gerber D., Hellmann S., Stadler C., Keyword-driven sparql query generation leveraging background knowledge, in: *Web Intelligence and Intelligent Agent Technology (WIIAT)*, Vol. 1, IEEE, pp. 203-210, 2011.

[9] Frantzi K., Ananiadou S., Mima H., Automatic recognition of multi-word terms: the c-value/nc-value method, *International Journal on Digital Libraries* 3 (2) 115, 2000.

[10] Daum S., Borrmann A., Langenhan C., Petzold F., Automated generation of building fingerprints using a spatio-semantic query language for building information models, eWork and eBusiness in *Architecture, Engineering and Construction*: ECPPM 2014 87, 2014.

[11] Langenhan C., Weber M., Liwicki M., Petzold F., Dengel A., Graph-based retrieval of building information models for supporting the early design stages, *Advanced Engineering Informatics* 27 (4) 413-42, 2013.

[12] Khalili A., Chua D., Ifc-based graph data model for topological queries onbuilding elements, *Journal of Computing in Civil Engineering* 0 (0) 04014046,

2013.

[13] Min H., Zhewen H., Ontology-driven tunnel construction information retrieval and extraction, in: *Control and Decision Conference* (2014 CCDC), The 26th Chinese, IEEE, 14, pp. 4741-4746, 2014.

[14] Zhang L., Issa R. R., Ontology-based partial building information model extraction, *Journal of Computing in Civil Engineering* 27 (6) 576-584, 2012.

[15] Nepal M. P., Staub-French S., Pottinger R., Zhang J., Ontology-based feature modeling for construction information extraction from a building information model, *Journal of Computing in Civil Engineering* 27 (5) 555-569, 2012

[16] Le T., Jeong H. D., Interlinking life-cycle data spaces to support decision making in highway asset management, *Automation in Construction* 64, pp. 54-64, 2016.

[17] Won J., Lee G., Cho C., No-schema algorithm for extracting a partial model from an ifc instance model, *Journal of Computing in Civil Engineering* 27 (6) 585-592, 2013.

[18] Wülfing A., Windisch R., Scherer R., A visual bim query language, eWork and eBusiness in *Architecture, Engineering and Construction*: ECPPM 2014 157, 2014.

[19] Zhang J., El-Gohary N., A semantic similarity-based method for semi-automated ifc extension, 2015.

[20] Lin J., Hu Z., Zhang J., Yu F., A natural-language-based approach to intelligent data retrieval and representation for cloud bim, *Computer-Aided Civil and Infrastructure Engineering*, 2015.

[21] Cambria E., White B., Jumping NLP curves: a review of natural language processing research [review article], *Computational Intelligence Magazine*, IEEE 9 (2) 48-57, 2014.

[22] Toutanova K., Klein D., Manning C. D., Singer Y., Feature-rich part-of-speech tagging with a cyclic dependency network, in: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*-Volume 1, Association for Computational Linguistics, pp. 173-180, 2003

[23] Cunningham H., Maynard D., Bontcheva K., Tablan V., Gate: an architecture for development of robust hlt applications, in: *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, pp. 168-175, 2002.

[24] Webster J. J., Kit C., Tokenization as the initial phase in nlp, in: *Proceedings of the 14th conference on Computational linguistics*-Volume 4, Association for Computational Linguistics, pp. 1106-1110, 1992.

[25] Zhao H., Kit C., Integrating unsupervised and supervised word segmentation: *The role of goodness measures, Information Sciences* 181 (1) 163-183, 2011.

[26] Lesk M., Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone, in: *Proceedings of the 5th annual international conference on Systems documentation*, ACM, pp. 24-26, 1986.

[27] Yarowsky D., Unsupervised word sense disambiguation rivaling supervised methods, in: *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics, pp. 189-196, 1995.

[28] Navigli R., Word sense disambiguation: A survey, *ACM Computing Surveys (CSUR)* 41 (2) 10, 2009.

[29] Marcus M., New trends in natural language processing: statistical natural language processing, *Proceedings of the National Academy of Sciences* 92 (22) 10052-10059, 1995.

[30] Costa-Jussa M. R., Farrús M., Mariño J. B., Fonollosa J. A., Study and comparison of rule-based and statistical catalan-spanish machine translation systems, *Computing and Informatics* 31 (2) 245-270, 2012.

[31] Erk K., Vector space models of word meaning and phrase meaning: A survey, *Language and Linguistics Compass* 6 (10) 635-653, 2012.

[32] Harris Z. S., Distributional structure, Word, 1954.

[33] Turney P. D., Pantel P., From frequency to meaning: Vector space models of semantics, *Journal of artificial intelligence research* 37 (1) 141-188, 2010.

[34] Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J., Distributed representations of words and phrases and their compositionality, in: *Advances in neural information processing systems*, pp. 3111-3119, 2013.

[35] Miller G. A., Wordnet: a lexical database for english, *Communications of the ACM* 38 (11) 39-41, 1995.

[36] Justeson J. S., Katz S. M., Technical terminology: some linguistic properties and an algorithm for identification in text, *Natural Language Engineering* 1 (01) 9-27, 1995.

[37] Mikolov T., Chen K., Corrado G., Dean J., Eficient estimation of word representations in vector space, *arXiv preprint arXiv:1301*.3781, 2013.

[38] Gale W. A., Church K. W., A program for aligning sentences in bilingual corpora, *Computational linguistics* 19 (1) 75-102, 1993.