INTERNATIONAL WORKSHOP

ON DEFINITION EXTRACTION

held in conjunction with the International Conference

RANLP - 2009, 14-16 September 2009, Borovets, Bulgaria

PROCEEDINGS

Edited by
Gerardo Sierra, Mara Pozzi and Juan-Manuel Torres

Borovets, Bulgaria

18 September 2009

International Workshop

ON DEFINITION EXTRACTION

PROCEEDINGS

Borovets, Bulgaria
18 September 2009

ISBN 978-954-452-013-7

Designed and Printed by INCOMA Ltd.
Shoumen, Bulgaria

Foreword

In the last few years the automatic extraction of definitions from textual data has become a common research topic in several domains of Natural Language Processing. These include:

- Definition extraction as a methodological resource for fields as different as computational semantics, information extraction, text mining, ontology development, WEB semantics and e-learning.
- The conception of definition extraction as a self-challenging task, in particular in computational lexicography and terminography, fields oriented towards the design and implementation of electronic tools such as lexical knowledge bases, machine-readable dictionaries, terminological databases, thesauri, machine translation systems or question-answering systems.

However, in contrast to the general use of definition extraction in multiple domains, there is no specific forum for sharing information about methodologies, tools, evaluation techniques or applications related to this field. Therefore, the goal of this workshop is to provide an opportunity to discuss theoretical and applied issues regarding definition extraction, such as:

- Contributions concerning the state of the art in definition extraction.
- Concrete applications of definition extraction in scientific or technical fields.
- The newest techniques to recognise and extract definitions candidates from running text using symbolic or statistical methods.
- Demonstration of computational tools for extracting definitions from large corpora.

The ten accepted papers report recent research initiatives on the topic of definition extraction and its applications.

The first paper, A formal scope on the relation between definitions and verbal predications by César Aguilar and Gerardo Sierra, outline a formal description of grammatical relations found in definitional contexts in Spanish and describe syntactic patterns relating definitions and predications and the usefulness of these patterns for the identification of definitions in technical corpora.

In the paper Description and evaluation of a definition extraction system for Spanish language, Rodrigo Alarcón, Gerardo Sierra and Carme Bach present a description and evaluation of a pattern-based approach for definition extraction in Spanish specialised texts based on the search for definitional verbal patterns related to analytical, extensional, functional and synonymical definitions.

Enriching a lexicographical tool with domain definitions: Problems and solutions by María Barrios, Guadalupe Aguado de Cea y José Ángel Ramos describes the problems faced by definition extraction methods due to poor definition construction and proposes some solutions.

In the paper Extraction of author's definitions using indexed reference identification, Marc Bertin, Iana Atanassova and Jean-Pierre Descles explore the establishment of relations between definitions and authors by using indexed references based on a linguistic ontology for the extraction of definitions from multilingual corpora of scientific texts.

The paper Evolutionary algorithms for definition extraction, by Claudia Borg, Mike Rosner and Gordon Pace, explores the use of machine learning methods to extract definitions. It reports the positive results obtained by the use of genetic programming and genetic algorithms to learn the relative importance of typical linguistic forms of definitions.

Language independent system for definition extraction: First results using learning algorithms, by Rosa Del Gaudio and António Branco, presents several language-independent approaches to deal with unbalanced data sets applied to two corpora in different languages for definition extraction using machine learning algorithms.

Gerard de Melo and Gerhard Weikum's paper Extracting Sense-Disambiguated Example Sentences From Parallel Corpora investigates to what extent sense-specific example sentences can be extracted from parallel corpora using lexical knowledge bases for multiple languages as a sense index to disambiguate word senses.

In her paper A proposal for a framework to evaluate feature relevance for terminographic definitions, Selja Seppälä proposes a theoretical and methodological terminology framework to evaluate relevant features obtained from definition extraction procedures for terminographical purposes.

In the paper Linguistic realization of conceptual features in terminographic dictionary definitions, Esperanza Valero Doménech and Amparo Alcina Caudet report the result of manual analysis of specialised dictionary definitions to identify relevant conceptual features and their linguistic realisation to extract and generate definitions.

Finally, Eline Westerhout's paper entitled Definition extraction using linguistic and structural features presents a promising approach to definition extraction in Dutch using a combination of linguistic (n-grams, type of article, type of noun) and structural information (layout, position).

We hope that this workshop will provide a forum for interaction among members of different research communities, a means for participants to increase their knowledge and understanding of the potential of definition extraction and a means for promoting definition extraction as a consolidated domain of NLP.

September 2009

Gerardo Sierra
María Pozzi
Juan-Manuel Torres

Chair

Gerardo Sierra, Universidad Nacional Autónoma de México, Mexico City, Mexico

Programme Committee

Teresa Cabré, Universitat Pompeu Fabra, Barcelona, Spain
Patrick Drouin, Université de Montréal, Montréal, Canada
Thierry Fontenelle, Microsoft Corporation, USA
Adam Kilgarriff, University of Sussex, Sussex, United Kingdom
John McNaught, National Center for Text Mining, Manchester, United Kingdom
Véronique Malaisé, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
Alfonso Medina, Universidad Nacional Autónoma de México, Mexico City, Mexico
Paola Monachesi, Universiteit Utrecht, Utrecht, The Netherlands
María Pozzi, El Colegio de México, Mexico City, Mexico
Paolo Rosso, Universitat Politcnica de València, Valencia, Spain
Juan-Manuel Torres, Université d'Avignon, Avignon, France

Organising Committee

César Aguilar, Universidad Autnoma de Quertaro, Queretaro, Mexico
Rodrigo Alarcón, Universidad Nacional Autónoma de México, Mexico City, Mexico
Carme Bach, Universitat Pompeu Fabra, Barcelona, Spain
Héctor Jiménez, Universidad Autónoma Metropolitana, Mexico City, Mexico
Horacio Saggion, University of Sheffield, Sheffield, United Kingdom

Table of Contents

Workshop Program

Friday, September 18, 2009

9:20–9:30 Welcome and opening Remarks

9:30–10:00 *Language independent system for definition extraction: first results using learning algorithms*
Rosa Del Gaudio and António Branco

10:00–10:30 *A Proposal for a framework to evaluate feature relevance for terminographic definitions*
Selja Seppälä

11:00–11:30 *Linguistic realization of conceptual features in terminographic dictionary definitions*
Esperanza Valero and Amparo Alcina

11:30–12:00 *A formal scope on the relations between definitions and verbal predications*
César Aguilar and Gerardo Sierra

12:00–12:30 *Extraction of author's definitions using indexed reference identification*
Marc Bertin, Iana Atanassova and Jean-Pierre Descles

14:00–14:30 *Evolutionary algorithms for definition extraction*
Claudia Borg, Mike Rosner and Gordon Pace

14:30–15:00 *Enriching a lexicographic tool with domain definitions: problems and solutions*
María A. Barrios, Guadalupe Aguado de Cea and José Ángel Ramos

15:00–15:30 *Description and evaluation of a pattern based approach for definition extraction*
Rodrigo Alarcón, Gerardo Sierra and Carme Bach

16:00–16:30 *Definition extraction using linguistic and structural features*
Eline Westerhout

16:30–17:00 *Extracting sense-disambiguated example sentences from parallel corpora*
Gerard de Melo and Gerhard Weikum

A formal scope on the relations between definitions and verbal predications

César Aguilar
Facultad de Lenguas y Letras
Universidad Autónoma de Querétaro
Centro Universitario, s/n C.P. 76010
Querétaro, México
Caguilar@iingen.unam.mx

Gerardo Sierra
Instituto de Ingeniería
Universidad Nacional Autónoma de México
Cubículo 3, Basamento, Torre de Ingeniería
C.U., C.P. 04510, México D.F.
GsierraM@iingen.unam.mx

Abstract

This paper outlines a formal description of grammatical relations between definitions and verbal predications found in Definitional Contexts in Spanish. It can be situated within the framework of Predication Theory, a model derived from Government & Binding Grammar. We use this model to describe: (i) the syntactic patterns that establish the relationship between definitions and predications; (ii) how useful these patterns are for the identification of definitions in technical corpora.

Keywords

Definition Extraction, Types of Definitions, Predication, Predicative Phrase.

1. Introduction

The (semi-)automatic recognition of terms and definitions in a corpus is an important task to research areas such as computational lexicography, terminology, language engineering and others. In the case of term recognition, several works report successful methodologies, computational tools and experiments that aim to identify and extract, in a no-supervised way, term candidates from large specialized corpora (e. g. Cabré, Estopà & Vivaldi 2001).

However, the automatic recognition of definitions presents a much higher degree of complexity, since definitions are linguistic structures used to formulate concepts (Sager, 1990). In contrast to terms, which are considered language units whose function is to refer specific entities in a scientific or technical knowledge domain, definitions condense information and establish several conceptual relations, with the purpose to delimitate the essential properties or attributes that characterize an entity in relation to others.

There are currently many authors that have proposed different methodologies for identifying candidates to definitions, considering both linguistic and statistical points of views. Some relevant methodologies are:

- Definitional Sentences (fr. *énonces définitoires*): Auger (1997), Rebeyrolle (2000).
- Terms in Contexts: Pearson (1998).
- Knowledge-Rich Contexts: Meyer (2001).
- Mining Definitions on Texts: Malaisé, Zweigenbaum & Bachimont (2005).

In accordance with these methodologies, in this paper we present a methodology to identify different types of definitions in technical corpora, considering that these definitions are configured as grammatical patterns, in particular, as phrase structures. These patterns are linked to verbal predications with syntactic regularities.

For the syntactic analysis of these patterns, we use a formal model called Predication Theory (henceforth, PredT). This model is formulated within the framework of Government & Binding Grammar (Rothstein, 1983; Bowers 1993, 2001. So, the PredT allows us to describe, in a formal way, the grammatical relations that definitions establish with verbal predications. Taking this relationship into account, it is possible to identify good candidates to definitions considering their association with verbal predications, specifically when these definitions are introduced in scientific and technical texts.

2. Definitional Contexts

We situate this analysis within the framework of Definitional Contexts (or DCs) extraction. According to Sierra *et al.* (2008), a DC is a discursive structure that contains relevant information to define a term. A DC has at least two constituents: a term and a definition, and usually linguistic or metalinguistic forms, such as verbal phrases, typographical markers and/or pragmatic patterns. An example is:

1. In general, the **paraprofessional workers** are defined as those persons who are engaged in the provision of social care or social services, but who do not have professional training or qualifications.

Workshop On Definition Extraction 2009 - Borovets, Bulgaria, pages 1–6,

According to this example, the term *Paraprofessional workers* is emphasised by the use of bold font; the verbal predication *are defined as* links the term *paraprofessional workers* to the actual definition *those persons who are engaged...* The term, the verbal predication and the definition are discursive units introduced by the pragmatic pattern *In general.* These are the three units that constitute the main syntactic sequence of a DC.

In this work we study this kind of DCs in Spanish, where the association between definitions and verbal predications is made explicit.

3. A formal description respect to predication

Taking into consideration that these sequences are composed of a verbal predication and definitions, several authors have found and reported such sequences for English (Pearson, 1998; Meyer 2001; Malaisé, Zweigenbaum & Bachimont, 2005) and French (Auger 1997, Rebeyrolle 2000). These authors have considered the use of these types of verbal predications as useful patterns for the (semi)automatic extraction of information associated to definitions.

However, none of these authors have analysed the nature of the relations between predications and definitions. In this paper, we focus on the description of their nature at the syntactic level, based on the PredT as a pertinent formal model for explaining the relations established between verbal predications and definitions.

3.1 Predication theory in GB grammar

Grosso modo, PredT is a model derived from Government & Binding Grammar, formulated by Chomsky (1981). PredT postulates that all predications indicate a semantic relationship between an entity and a particular property or characteristic feature. Syntactically, PredT explains all verbal predications as a type of phrase, structured around a relation *X-is-a-Subject-of/Y-is-a-predicate-of*. This relation is regulated by a syntactic rule named *rule of predicate linking*, proposed by Rothstein (1983). Examples of these relations are:

2. a. John is an intelligent professor.
   b. John considers his father as an intelligent professor.

Following Rothstein's explanation, Bowers (1993, 2001) develops a simple model to describe the syntactic configuration of these phrases, called Predicative Phrase (PrP). The PrP is mapped by a non-lexical head (that is, a functional head), and its grammatical behaviour is similar to that of phrases such as Inflexional Phrase (IP) or Complement Phrase (CP). A graphical tree representation of a PrP is:

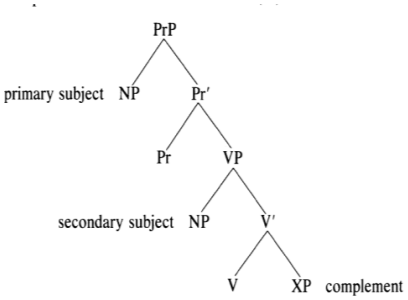


Figure 1: Tree representation for PrP, according to Bowers (1993: 596).

Figure 1 shows the basic elements that make up a PrP. Bowers recognise a functional head with the features $^{+/-}$ predicative (Pr). This head maps two Subjects, a primary subject in the position of Specifier of PrP (represented by a Noun Phrase or NP); and a secondary subject, in the position of Specifier of Verbal Phrase or VP (often a NP). Finally, both subjects, the VP and the PrP are linked to one or several complements, which assume many phrasal representations (e.g. NP, IP, CP, and so on).

3.2 Primary and secondary predications

Based on this distinction between primary and secondary subjects, it is possible to recognise two types of predications:

- Simple or primary predication, consisting of a subject to the left of the verb (in position of Specifier of PrP), and a predicate to the right of the verb. An example in Spanish is:

3. [Una computadora [es [un tipo de máquina electrónica que sirve para hacer operaciones $_{PrP}$] $_{VP}$] $_{IP}$] (Eng. [A computer [is [a kind of electronic machine used to make operations $_{PrP}$] $_{VP}$] $_{IP}$]).

- Double or secondary predication, which integrates a primary subject in a pre-verbal position, a secondary subject (situated as Specifier of VP), and the predicate. For example, again, in Spanish:

4. [Turing [define una computadora [como un mecanismo electrónico que procesa conjuntos de datos $_{PrP}$] $_{VP}$] $_{IP}$] (Eng. [Turing [defines a computer [as a kind of electronic device that processes a set of data $_{PrP}$] $_{VP}$] $_{IP}$]).

In (4), the predicate *como un mecanismo electrónico*... (Engl. *as a kind of electronic device*...) affects the secondary subject *una computadora* (Engl. *a computer*), in accordance with the explanation provided by Bowers (1993). For our analysis, we consider both types of predications as regular patterns that syntactically codify sequences of terms, verbal predications and definitions.

4. Combinatory of patterns in DCs

Based on our formal description of PrP, it is possible to identify two types of patterns that structure particular sequences in DCs:

- In the case of primary predications, it codifies a sequence composed of a Term, a Verbal Predication and a Definition.
- In the case of secondary predication, it codifies a sequence composed of a specific Author, a Term, a Verbal Predication and a Definition.

4.1. Term + Verbal Predication + Definition

This sequence is a good example of a formulation of canonical definitional patterns, because the primary predication links directly a subject represented by a term, with a specific set of attributes codified in the PrP. These patterns are shown in the following sequences:

5. a. [El contenedor refrigerado $_{Term}$] [es $_{Verbal\ Predication}$] [una forma especializada de transporte de perecederos $_{Definition}$] (Eng. [The refrigerated container $_{Term}$] [is $_{Verbal\ Predication}$] [a specialized form to transport perishable goods $_{Definition}$])
   b. [Un esquema XML $_{Term}$] [representa $_{Verbal\ Predication}$] [el significado y la estructura de la información recibida desde una aplicación $_{Definition}$] (Eng. [An XML schema $_{Term}$] [represents $_{Verbal\ Predication}$] [the meaning and structure of the information received from an application $_{Definition}$]).
   c. [Una jerarquía de dependencias $_{Term}$] [se refiere a $_{Verbal\ Predication}$] [todas las tablas que incluyen referencias mutuas $_{Definition}$] (Eng. [A hierarchy of units $_{Term}$] [refers to $_{Term}$] [all tables that include references to each other $_{Definition}$]).

The sequence Term + Verbal Predication + Definition in cases 5 a-c is equivalent to the structure of primary predication. Therefore, the Term is situated in the position of Primary Subject, the Verbal Predication has the role of head of a VP, and the Definition is introduced through a PrP.

4.2. Author + Term + Verbal Predication + Definition

The second sequence we report here shows the sequence Author + Term + Verbal Predication + Definition. The characteristic feature of this pattern is that it explicitly points out the author (or authors) of the definition. This feature maps a semantic role, according to FrameNet (Baker, Fillmore and Lowe, 1998), concretely the author can be conceived as a Cognizer that associates certain Categories (the Definition) to a particular Item (that is, the Term). This is illustrated in the following examples:

6. a. [Carlos Godino $_{Author}$] [define $_{Verbal\ Predication}$] [la arquitectura naval $_{Term}$] [como la ciencia que se enfoca en la construcción de los buques $_{Definition}$] (Eng. [Carlos Godino $_{Author}$] [defines $_{Verbal\ Predication}$] [naval architecture $_{Term}$] [as the science that focuses on the construction of ships $_{Definition}$])
   b. [El artículo $_{Author}$] [describe $_{Verbal\ Predication}$] [la evolución de ecología del paisaje $_{Term}$] [como una ciencia integrativa y transdisciplinaria $_{Definition}$] (Eng. [The article $_{Author}$] [describes $_{Verbal\ Predication}$] [the evolution of landscape ecology $_{Term}$] [as an integrative and interdisciplinary science $_{Definition}$]).
   c. [Ø Podemos $_{Author}$] [considerar $_{Verbal\ Predication}$] [las computadoras programables modernas $_{Term}$] [como la evolución de sistemas antiguos de cálculo o de ordenación $_{Definition}$] (Eng. [We $_{Author}$] [can consider $_{Verbal\ Predication}$] [the modern programmable computers $_{Term}$] [as the evolution of ancient systems of calculation and management $_{Definition}$]).

Hence, the pattern followed by this sequence clearly refers to the author of a definition, as shown in 6 a-c. However, a syntactic behaviour observed in this pattern is its recurrent configuration in non-personal forms, i.e. impersonal and passive forms, for example:

7. a. [Se conoce como $_{Verbal\ Predication}$] [reenganche rápido $_{Term}$] [a la operación de cierre de un interruptor después de una falla $_{Definition}$] (Eng. [It is known as $_{Verbal\ Predication}$] [Quick Re-closing $_{Term}$] [to the operation of a switch after a fault $_{Definition}$]).
   b. [Los niveles relativos de los alcances de ola $_{Term}$] [fueron descritos como $_{Verbal\ Predication}$] [una función del parámetro de similitud de oleaje $_{Definition}$] (Eng. [The relative levels of the wave reach $_{Term}$] [were described as $_{Verbal\ Predication}$] [a function of the wave similarity parameter $_{Definition}$]

In these examples, we observe the use of non-personal verbal patterns as in (7a), where the clitic *Se* is inserted (Eng. *It*) to make the sentence impersonal, or in (7b), which is in the passive form. So, when these sequences assume a non-personal pattern, they become equivalent to primary predications, where there is not an explicit mention of the author of a definition.

5. Types of definitions linked to predications

Another aspect that we found in the relation between predications and definitions is the influence of the predication on the selection of a particular type of definition. In fact, this influence is important because we can establish and formalise a possible grammar model that helps to identify different kinds of definitions, given a primary or secondary predication.

Following Sierra *et al.* (2008) and Aguilar (2009), we outline a typology with 4 types of definitions: analytical, synonymical, functional and extensional. These definitions are derived from Aristotle's model:

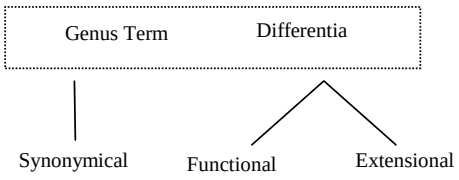


Figure 2. Typology of definitions based on Aristotle's model (Sierra *et al.* 2008: 81).

Figure 2 illustrates how these four types of definitions can be identified according to the presence/absence of Genus Term and/or Differentia in a good candidate of definition. So, when both the Genus Term and the Differentia are explicit, we have an analytical definition, which can be associated with two kinds of predication: primary predication or secondary predication.

When only the Genus Term is explicit, there is a certain relation of conceptual equivalence between the term and its definition. So, following Cruse (1986), we characterise this definition as synonymical.

In contrast, when only the Differentia is explicit, we have two options:

- The first one describes the use or the function of an entity represented by the term, that is, a functional definition.
- The second one enumerates all the components of a possible entity or a possible set, that is, an extensional definition.

In the following sections we briefly describe each type of definition.

5.1. Analytical definitions

This definition occurs associated with primary and secondary predications. In the case of primary predication, the analytical definition is integrated in a sequence Term + Verbal Predication + Definition. This definition does not explicitly state the author of a definition. For example:

8. [El apartarrayos $_{\text{Term}}$] [es $_{\text{Verbal Predication}}$] [un dispositivo $_{\text{Genus Term}}$] [que protege las instalaciones contra sobretensiones de origen atmosférico $_{\text{Differentia}}$] (Engl. [The lightning conductor $_{\text{Term}}$] [is $_{\text{Verbal Predication}}$] [a device $_{\text{Genus Term}}$] [that protects electrical systems against surges of atmospheric origin $_{\text{Differentia}}$]).

We propose a possible grammatical description model for this relation:

Definition	Genus Term	Differentia
Analytical (Simple Predication)	Noun Phrase = Noun + {Adjective Phrase/Prepositional Phrase}*	Complement Phrase = Relative Pronoun + Inflexional Phrase
		Prepositional Phrase = Preposition + Noun Phrase
		Adjective Phrase = Adjective + Noun Phrase

Table 1. Construction patterns derived from the relation between primary predication and analytical definition

In the case of secondary predications linked to analytical definitions, they follow the sequence Author + Term + Verbal Predication + Definition, where the Author is equivalent to the primary subject, the Term assumes the position of secondary subject, and the definition is introduced after the Verbal Predication. In this case, the adverbial particle *como* (Eng. *as/like*), or the preposition *por* (Eng. *for/by*) indicate the place of the definition:

Definition	Adverb/ Preposition	Genus Term	Differentia
Analytical (Secondary Predication)	*Como Por*	Noun Phrase = Noun + {Adjective Phrase/Prepositional Phrase}*	Complement Phrase = Relative Pronoun + Inflexional Phrase
			Prepositional Phrase = Preposition + Noun Phrase
			Adjective Phrase = Adjective + Noun Phrase

Table 2. Construction patterns derived from the relation between secondary predication and analytical definition

5.2. Synonymous definitions

The synonymous definitions have a syntactic relation with primary predications, specifically with the Genus Term, but not with the differentia. An example is:

9. [La tensión de base $_{\text{Term}}$] [se le llama también $_{\text{Verbal Predication}}$] [tensión unidad $_{\text{Genus Term}}$]. (Engl. [The base tension¨ $_{\text{Term}}$] [it is also called $_{\text{Verbal Predication}}$] [unit tension $_{\text{Genus Term}}$]).

In (9), we observe that the Term *la tensión de base* (Engl. *the base tension*) establishes a relation of cognitive equivalent with the Genus Term *tensión unidad* (Engl. *unit tension*). We formalise this relation in table 3:

Definition	Term	Genus Term
Synonymical (Primary Predication)	Noun Phrase = Noun + {Adjective Phrase/Prepositional Phrase}*	Noun Phrase = Noun + {Adjective Phrase/Prepositional Phrase}*

Table 3. Construction patterns derived from the relation between primary predication and synonymous definition

5.3. Functional definitions

The functional verbal pattern introduces a type of definition where the Genus Term is absent, but introduces a Differentia that describes the function or the use of a particular entity. The verbal pattern is also associated with a primary predication. The example is:

10. [La técnica de velocimetría de imágenes $_{\text{Term}}$] [permite $_{\text{Verbal Predication}}$] [medir la velocidad de un campo de flujo bi o tri dimensional $_{\text{Differentia}}$] (Engl. [The method of image velocimetry $_{\text{Term}}$] [allows $_{\text{Verbal Predication}}$] [to measure the speed of a flow field in two or three dimensions $_{\text{Differentia}}$]).

The formal description of this relation between predication and definition is:

Definition	Differentia
Functional (Primary Predication)	Infinitive Verb + Complement Phrase = Relative Pronoun + Inflexional Phrase + {Prepositional Phrase/Adjetive Phrase/Adverbial Phrase/Complement Phrase}*
	Infinitive Verb + Preposition + {Inflexional Phrase/Complement Phrase}*
	Prepositional Phrase = Preposición + Noun Phrase + {Prepositional Phrase/Adjetive Phrase/Adverbial Phrase/Complement Phrase}*
	Noun Phrase = Noun + {Prepositional Phrase/Adjetive Phrase/Adverbial Phrase/Complement Phrase}*

Table 4. Construction patterns derived from the relation between primary predication and functional definition

5.4. Extensional definitions

Finally, extensional definitions provide a complete list of the parts,,components or elements of a entity or set. In a similar way to functional definitions, extensional definitions are structured around a primary predication. An example is:

11. [La zona límite $_{\text{Term}}$] [incluye $_{\text{Verbal Predication}}$] [planicies costeras, marismas, áreas de inundación, playas, dunas y corales $_{\text{Differentia}}$] (Eng. [The border zone $_{\text{Term}}$] [includes $_{\text{Verbal Predication}}$][coastal plains, salt marshes, flood areas, beaches, dunes and corals $_{\text{Differentia}}$]).

Our syntactic description of this pattern is:

Definition	Preposition	Differentia
Extensional (Primary Predication)	Con (Eng With) De (Eng. Of)	Noun Phrase = Noun + {Adjective Phrase/Prepositional Phrase}*

Table 5. Construction patterns derived from the relation between primary predication and extensional definition

We can summarise all these patterns in table 6, considering some recurrent verbs in the position of head of PrP. These verbs are not exclusive, but their recurrence has been reported by Sierra *et al.* (2008), and Aguilar (2009):

Definition	Verbs	Associated Particles
Analytical (Primary Predication)	*referir* (to refer to) *representar* (to represent) *ser* (to be) *significar* (to signify/to mean)	*a* = to (preposition) (in the case of *referir*, it is a phrasal verb that inserts obligatory the preposition *a*)
Analytical (Secondary Predication)	*caracterizar* (to characterise) *comprender* (to include) *concebir* (to conceive) *conocer* (to know) *considerar* (to consider) *definir* (to define) *describir* (to describe) *entender* (to understand) *identificar* (to identify) *visualizar* (to visualise)	*como* = as/like (adverb) *por* = for/by (preposition)
Synonymy	*equivaler* (to be equivalent to) *llamar* (to call) *nombrar* (to name) *ser _ igual* (to be equal to) *ser _ similar* (to be similar to)	*también* = also (adverb) a = to (preposition) *igual a* = equal to (adverb phrase) *similar a* = similar to (adverb phrase)
Functional (Primary Predication)	*emplearse* (to employ + clicit "se") *encargar* (to be in charge of) *funcionar* (to function) *ocupar* (to occupy) *permitir* (to allow) *servir* (to serve) *usar* (to use) *utilizar* (to utilise / to use)	*de* = of (preposition) *para* = for (preposition)
Extensional (Primary Predication)	*componer* (to be composed of) *comprender* (to include) *consistir* (to consist of) *constar* (to consist of) *contar* (to have) *constituir* (to constitute) *contener* (to contain) incluir (to include) integrar (to integrate)	*de* = of (preposition) *por* = for/by (preposition) *con* = with (preposition)

Table 6. Verbs associated with definitions

6. Commentaries and conclusions

In this paper, we have outlined a formal description of the grammatical relations that can be established between definitions and verbal predications in DCs. We consider this is a pertinent way to analyse the syntactic behaviour of definitions in specialised texts, specifically when these definitions are linked to verbal predications.

We have described these verbal predications according to the PredT, a grammatical model useful to formalise patterns generated by the association of verbal predications to specific definitions. This description allowed us to distinguish:

- Two types of verbal predications: primary and secondary predications. Both predications entail particular types of definitions, depending on the verb that functions as the head of the predication.
- These predications play an important role in the selection and introduction of specific types of

definitions. In this paper, we have proposed a possible typology of definitions, based on the role played by predications. This typology considers four types of definitions: analytical, synonymical, functional and extensional.

- In addition, it is possible to observe that the relation established between the types of definitions with primary/secondary predications configure two sequences that structure two different kinds of DCs: (i) a sequence, Term + Verbal Predication + Definition, configured in primary predications which can be linked to analytical, synonymical, functional and extensional definitions and; (ii) another sequence, Author + Term + Verbal Predication + Definition, delineated by secondary predications which can be associated to secondary predications.

We think that the use of these patterns proposed in our analysis can sketch a useful grammatical model, applied to the task of (semi)automatic recognition and extraction of definitions in Spanish, from text corpora.

7. Acknowledgements

This paper was made possible by the financial support of the Consejo Nacional de Ciencia y Tecnología, CONACYT, and DGAPA-UNAM.

8. References

[1] C. Aguilar. Análisis lingüístico de definiciones en contextos definitorios. Ph. D. Thesis, National Autonomous University of Mexico, Mexico City, 2009.

[2] A. Auger. Repérage des énonces d'intérêt définitoire dans les bases de données textuelles. Thèse de Doctorat, Neuchâtel, Université de Neuchâtel, 1997.

[3] C. Baker, Ch. Fillmore, & J. Lowe. The Berkeley FrameNet Project. In Proceedings of the COLING-ACL, Montreal, Canada, 1998.

[4] J. Bowers. The Syntax of Predication. Linguistic Inquiry, 24(4): 591-636, 1993.

[5] J. Bowers. Predication. In Baltin, M. & C. Collins (eds.). The Handbook of Contemporary Syntactic Theory. Oxford, Blackwell: 299-333, 2001.

[6] T. Cabré, R. Estopà & J. Vivaldi. Automatic term detection. In Bourigault D., C. Jaquemin & M.C. L'Homme (eds.). Recent Advances in Computational Terminology. Amsterdam/Philadelphia, John Benjamins: 53-87, 2001.

[7] N. Chomsky. Lectures on Government and Binding. The Hague, Mouton de Gruyter, 1981.

[8] D. Cruse. Lexical Semantics. Cambridge, Cambridge University Press, 1986.

[9] V. Malaisé, P. Zweigenbaum & B. Bachimont. Mining defining contexts to help structuring differential ontologies. Terminology 11(1): 21-53, 2005.

[10] I. Meyer. Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. In Bourigault D., C. Jaquemin & M.C. L'Homme (eds.). Recent Advances in Computational Terminology, Amsterdam/Philadelphia, John Benjamins: 279-302, 2001.

[11] J. Pearson. Terms in Contexts. Amsterdam/Philadelphia, John Benjamins, 1998.

[12] J. Rebeyrolle. Forme et fonction de la définition en discours, Thèse de Doctorat. Toulouse, Université Toulouse-Le Mirail, 2000.

[13] S. Rothstein. The Syntax Forms of Predication, Ph. D. Thesis. Cambridge, Mass., MIT, 1983.

[14] J.C. Sager. Essays on Definitions, Amsterdam/Philadelphia, John Benjamins., 1990.

[15] G.Sierra, R. Alarcon, C. Aguilar & C. Bach. Definitional Verbal Patterns for Semantic Relation Extraction. In Auger A. y C. Barrière (eds.). Pattern-based Approaches to Semantic Relation Extraction. Special issue of *Terminology*, 14(1): 74–98, 2008.

Description and Evaluation of a Definition Extraction System for Spanish language

Rodrigo Alarcón
Universidad Nacional Autónoma de México, Grupo de Ingeniería Lingüística
Torre de Ingeniería, Basamento 3, Mexico City
ralarconm@iingen.unam.mx

Gerardo Sierra
Universidad Nacional Autónoma de México, Grupo de Ingeniería Lingüística,
Torre de Ingeniería, Basamento 3, Mexico City
gsierram@iingen.unam.mx

Carme Bach
Universitat Pompeu Fabra, Grupo IULATERM, Departament de Traducció i Ciències del Llenguatge,
Roc Boronat, 138, Barcelona, Spain
carme.bach@upf.edu

Abstract

In this paper we present a description and evaluation of a pattern-based approach for definition extraction in Spanish specialised texts. The system is based on the search for definitional verbal patterns related to four different kinds of definitions: analytical, extensional, functional and synonimical. This system could be helpful in the development of ontologies, databases of lexical knowledge, glossaries or specialised dictionaries.

Keywords

Definition extraction, definitional contexts, definitional verbal patterns, pattern-based approach.

1. Introduction

There is a growing interest in the development of systems for the automatic extraction of information that describe the meaning of terms. This information occurs in structures commonly called *definitional contexts* (DCs), which are structured by a series of lexical and metalinguistic patterns that can be automatically recognised. In this context, in this paper we present a work focused on developing a system for the automatic extraction of definitional contexts on Spanish language specialised texts. This system looks for instances of definitional verbal patterns, filters non-relevant contexts, identifies the main constituent elements on the candidates, i.e., terms and definitions, and performs an automatic ranking of the results.

Firstly, we will describe the structure of DCs; secondly, we provide a short review of related works; we then present the methodology followed for the automatic extraction of DCs together with an evaluation of this methodology; and lastly, we propose some future work.

2. Definitional Contexts in Specialised Texts

A definitional context is a textual fragment from a specialised text where a definition of a term is given. Its basic structure consists of a term (T) and its definition (D), both elements being connected by typographic or syntactic patterns. Typographic patterns are punctuation marks (comas, parenthesis), while syntactic patterns include definitional verbs –such as *definir* (to define) or *significar* (to signify)– as well as discursive markers –such as *es decir* (that is, lit. (it) is to say), or *o sea* (that is, lit. or be-subjunctive)–. Apart from these, DCs can include pragmatic patterns (PPR), which provide conditions for the use of the term or clarify its meaning, as in *en términos generales* (in general terms) or *en este sentido* (in this sense). For example:

"Desde un punto de vista práctico, los opioides se definen como compuestos de acción directa, cuyos efectos se ven antagonizados estereoespecíficamente por la naloxona."

In this case, the term *opioides* is connected to its definition (*compuestos de acción directa […]*) by the verbal pattern *se definen como* (are defined as), while the general sense of the context is modified by the pragmatic pattern *desde un punto de vista práctico* (from a practical point of view).

3. Advances in Definitional Contexts Extraction

Definition extraction from specialised texts has become a relevant task in the field of information extraction. In order to extract definitional information, the most common strategy is to extract certain recurrent patterns, which are commonly found in DCs.

The use of this kind of patterns has been applied on different scenarios. One of the first descriptive works can be found in [1], in which the behaviour of the contexts where terms occur is described. This work states that, when authors define a term, they usually employ typographic patterns to visually highlight the presence of terms and/or definitions, as well as lexical and metalinguistic patterns connecting DCs elements by means of syntactic structures. [2] reinforces this idea was reinforced and also explained the fact that definitional patterns can provide keys for the identification of the type of definition occurring in DCs, which facilitates the task of ontology development.

Workshop On Definition Extraction 2009 - Borovets, Bulgaria, pages 7–13,

Regarding applied works, [3] reports a system called *Definder* for the automatic extraction of definitions from medical texts in English. In the same line of research, other works have been focused on DCs extraction from specialised texts in other languages, for example German [4], Portuguese [5] or Spanish [6]. Definition extraction has also been used as a previous step for the automatic extraction of semantic relations or the automatic development of ontologies [7], [8], as well as for obtaining knowledge for the development of *eLearning* technologies [9].

Furthermore, the automatic extraction of definitions has been focused on direct Web exploitation. That is the case of the work reported in [10] whose main goal is the extraction of definitions from on on-line sources for question answering systems. [11] reports an application called *GlossExtractor*, that works on the Web, mainly online glossaries and Web specialised documents, also for the automatic extraction of definitions, but starting from a list of predefined terms. [12] developed a system called *DefExplorer* for definition extraction of Web documents for the Chinese Language.

All of these systems start from the search of specific definitional patterns in each language and they also integrate procedures for filtering non-relevant contexts, i.e., contexts that contain a definitional pattern that does not yield an actual definitional context. Finally, all of these methodologies are based on the exploitation of specialised documents, being the direct Web exploitation a recently incorporated process.

4. ECODE

As we have mentioned before, the main purpose of a definitional context extractor is to simplify the search of relevant information about terms, by means of searching for occurrences of definitional patterns.

An extractor that only retrieves those occurrences of definitional patterns would be a useful system for terminographical work. However, the manual analysis of the retrieved occurrences would still imply an effort that could be simplified by an extractor that includes the automatic processing of the obtained information.

Therefore, we propose a methodology that includes not only the extraction of occurrences of definitional patterns, but also a filtering process of non-relevant contexts (i.e. non definitional contexts), the automatic identification of the possible constitutive elements of a DC: terms and definitions, and a final automatic ranking of the results. This system is called *ECODE: extractor de contextos definitorios* (definitional contexts extractor).

A general overview of the system is shown in figure 1. It can be seen that the system input consists of a corpus tagged with POS categories, since some of them are necessary in the different processes of the system. It can also be seen that the main three processes are: a) the extraction of DC candidates, b) the analysis of DC candidates, and c) the evaluation of DC candidates.

The extraction of DC candidates is a process that uses a grammar of verbal patterns with some specific parameters: the definitional verbs to search for and the nexus that can also be part of the pattern, i.e., the adverb *como* (as) in the pattern *se define como* (it is defined as). In this case, the grammar shall also include constraints on the verbal times and grammatical person in which each verb can occur, as well as the different positions for each verb where the term can occur in a DC.

Once the DC candidates are extracted, they are analysed in the next process, which is carried out in two steps: the filtering of non-relevant candidates, and the identification of their constituent elements. The filtering process makes use of a set of linguistic and contextual rules to determine those cases where no DCs are found, while the identification of their constituent elements makes use of a decision tree, which also analyses the grammar of verbal patterns in order to identify the term and its definition on each DC candidate.

Finally, the system performs an automatic ranking of the candidates proposed as DCs. This process use a set of heuristic rules and aims to identify those candidates that follow a prototypical structure of terms and definitions.

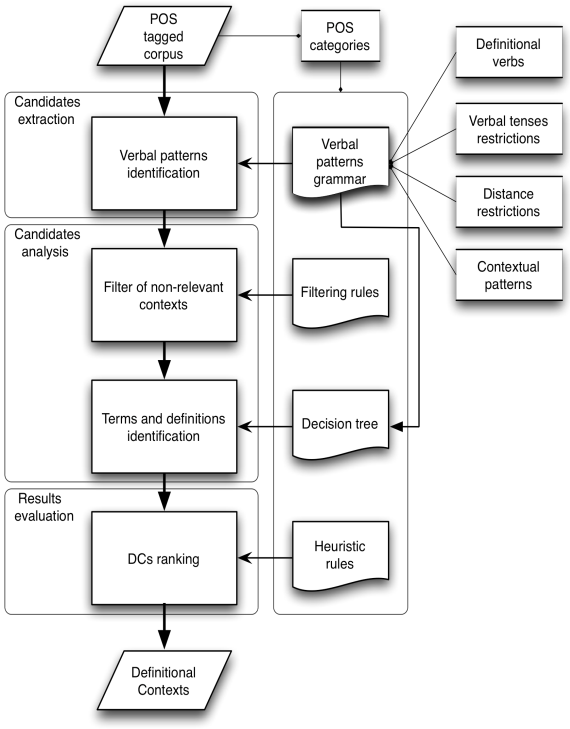


Fig. 1. System architecture.

4.1 Candidates extraction

The ECODE was developed taking the IULA's Technical Corpus from the Institut Universitari de Lingüística Aplicada (UPF) as starting point. This corpus consists of specialised documents in the fields of Law, Genome, Economy, Environment, Medicine, Informatics and General Language. First, we manually developed a grammar of verbal patterns for Spanish. We identified 29 verbs related to four different types of definitions: analytical, extensional, functional and synonimical. The whole set of verbal patterns is shown in table 1.

Table 1. Definitional Verbal patterns

Analytical verbal patterns
ser + artículo (to be + article)
consistir en (to consist in)
caracterizar como/por (to characterize as/for)
concebir como (to conceibe as)
considerar como (to consider as)
describir como (to describe as)
comprender como (to understand as)
definir como (to define as)
entender como (to understand as)
conocer como (to known as)
denominar como/∅ (to denominate as/∅)
llamar como/∅ (to call as/∅)
nombrar como/∅ (to name as/∅)
Extensional verbal patterns
comprender (to comprehend)
contener (to contain)
incluir (to include)
integrar (to integrate)
constar de (to comprise of)
contar de/con (to count of/with)
consistir de/en (to consist of/in)
formar de/por (to form of/by)
componer de/por (to compose of/by)
constituir de/por (to constitute of/by)
Functional verbal patterns
permitir (to allow)
encargar de (to undertake of)
funcionar como/para (to function as/for)
ocupar como/para (to occupy as/for)
servir como/en/para (to serve as/in/for)
usar como/en/para (to use as/in/for)
emplear como/en/para (to employ as/in/for)
utilizar como/en/para (to utilise as/in/for)
Synonimical verbal patterns
conocer también (to known also)
denominar (to denominate also)
llamar (to call also)
nombrar (to name also)

From the table above, we can see different verbs associated to different types of definitions. In some cases, the verbs can occur together with different grammatical particles and can be associated with more than one type of definition, such as the verb *denominar* (to denominate), which can occur in analytical or synonimical DCs with the nexus *como* (as) or *también* (also), respectively.

The verbal patterns were searched for taking into account the next constraints:

Verbal forms: infinitive, participle and conjugate forms.

Verbal tenses: present and past for verbs without nexus, any verbal tense for verbs with nexus.

Person: 3rd person singular and plural for verbs without nexus, any for verbs with nexus.

Distance: each nexus was searched for within a distance of 15 possible words.

With these restrictions, the system obtains a set of DC candidates that are next annotated with *contextual tags*. These simple tags function as borders in the next automatic processes. For each occurrence, the definitional verbal pattern was annotated with "<dvp></dvp>"; everything after the pattern with "<left></left>"; everything before the pattern with "<right></right>"; and finally, in those cases where the verbal pattern includes a nexus, like the adverb *como* (as), everything between the verbal pattern and the nexus was annotated with <nexus></nexus>. Here is an example of a DC annotated with contextual tags:

<left>El metabolismo</left> <dvp>puede definirse </dvp> <nexus>en términos generales como</nexus> <right>la suma de todos los procesos químicos (y físicos) implicados.</right>

4.2 Candidates analysis

Once the DCs were extracted and annotated with definitional verbal patterns they were analysed with the purpose of filtering non-relevant contexts. We applied this step based on the fact that definitional patterns are used not only in definitional sentences but also in a wider range of sentences. In the case of verbal patterns, some verbs tend to have a higher metalinguistic meaning than others. That is the case of *definir* (to define) or *denominar* (to denominate), vs. *concebir* (to conceive) or *identificar* (to identify), where the last two are used in different contexts. Moreover, verbs having a high metalinguistic meaning are not used only for defining terms.

To develop this process, a manual analysis was carried out to determine the type of grammatical particles or syntactic sequences occurring in those cases where a DVP was not used to define a term.

These particles and syntactic sequences were found in some specific positions, for example: negation particles such as *no* (not) or *tampoco* (either) were found in the first

position before or after the DVP; adverbs like *tan* (so), *poco* (few) as well as sequences like *poco más* (not more than) were found between the definitional verb and the nexus *como*; also, syntactic sequences such as adjective + verb were found in the first position after the definitional verb.

Thus, taking this and other frequently combinations into consideration as well as the contextual tags previously annotated, the systems filters contexts as shown in the following examples:

Rule: **NO <left>**

<left>En segundo lugar, tras el tratamiento eficaz de los cambios patológicos en un órgano pueden surgir problemas inesperados en tejidos que previamente **no </left>** <dvp>se identificaron</dvp> <nexus> como </nexus> <right> implicados clínicamente, ya que los pacientes no sobreviven lo suficiente.</right>

Rule: **<nexus> CONJUGATED VERB**

<left>Ciertamente esta observación tiene una mayor fuerza cuando el número de categorías </left> <dvp> definidas</dvp> **<nexus> es** pequeño como</nexus> <der>en nuestro análisis.</der>.

Once the non-relevant contexts were filtered, the next process was the identification of terms and definitions in the DC candidates. Depending on each DVP, the terms and definitions may appear in some specific positions in Spanish DCs. For example, in DCs containing the verb *definir* (to define), the term may occur in left, nexus or right position (T *se define como* D; *se define* T *como* D; *se define como* T D), while in DCs containing the verb *significar* (to signify), terms may appear only in left position (T *significa* D). Therefore, in this phase the automatic process is highly related to deciding the positions in which the constituent elements could appear.

We decided to use a decision tree to solve this problem, i.e., to detect by means of logic inferences the probable positions of terms, definitions and pragmatic patterns. We established some simple regular expressions to represent each constituent element[1]:

T = BRD (Det) + N + Adj. {0,2} .* BRD

PPR = BRD (sign) (Prep | Adv) .* (sign) BRD

D = BRD (Det) + N

As in the filtering process, the contextual tags function as borders to demarcate decision tree's instructions. In addition, each regular expression could function as a border. At the first level, the branches of the tree correspond to the different positions in which constituent elements may occur (left, nexus or right). At the second level, the branches correspond to the regular expressions of each DC element. The nodes (branches conjunctions) correspond to decisions taken from the attributes of each branch and are also horizontally related by *If* or *If Not* inferences, and vertically through *Then* inferences. Finally, the leaves correspond to the assigned position of each constituent element.

Hence, figure 2 shows an example of the decision tree inferences needed to identify constituent elements[2] in left position:

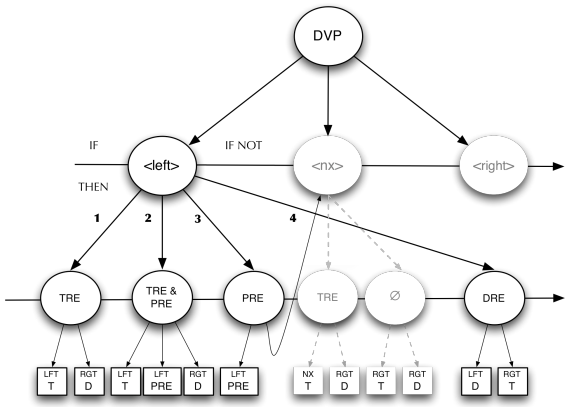


Fig. 2. Example of the identification of DCs elements.

This tree should be interpreted as follows: Given a series of DVPs occurrences:

1. *If* left position corresponds *only* to a term regular expression, *then*:

<left> = term | <right> = definition.

If Not:

2. *If* left position corresponds to a term regular expression and a pragmatic pattern regular expression, *then*:

<left> = term & pragmatic pattern | <right> = definition.

If Not:

3. *If* left position *only* corresponds to a pragmatic pattern regular expression, *then*[3]:

<left> = pragmatic pattern | *If* nexus corresponds *only* to a term regular expression, *then* <nexus> = term & <right> = definition; *If Not* <right> = term & definition.

4. *If* left position corresponds only to a definition regular expression, *then*:

[1] Where: Det= determiner, N= name, Adj= adjective, Prep= preposition, Adv= adverb, BRD= border and ".*"= any word or group of words.

[2] TRE = term regular expression | PRE = pragmatic pattern regular expression | DRE = definition regular expression.

[3] In some cases the tree must resort to other position inferences to find terms and definitions.

<left> = definition | <right> = term.

To exemplify this we can observe the next context:

"<left>En sus comienzos</left> <dvp>se definió</dvp> <nexus>la psicología como </nexus><right>"la descripción y la explicación de los estados de conciencia" (Ladd, 1887).</right>"

Once the DVP was identified as a CDVP – *definir como* (to define as) – the tree infers that left position:

1. Does not correspond only to a TRE.

2. Does not correspond to a TRE and a PRE.

3. It corresponds only to a PRE.

Then: left position is a pragmatic pattern (*En sus comienzos*). To identify the term and definition the tree goes to nexus's inferences and finds that:

1. It does correspond only to a TRE.

Then: nexus's position corresponds to the term (*la psicología*) and right's position corresponds to the definition ("la descripción y la explicación de los estados de conciencia […]").

As a result, the processed context was reorganised into terminological entries, as in the following example:

Table 2. Example of ECODE results

TERM	Psicología
DEFINITION	"la descripción y la explicación de los estados de conciencia" (Ladd, 1887)
PRAGMATIC PATTERN	En sus comienzos
VERBAL PATTERN	se definió como

4.3 Evaluation of results

In order to complement the system's processes described above, we decided to include an automatic ranking of the results. This automatic evaluation aims to identify those contexts with more prototypical structures of terms and definitions as well as structures reinforced by typographic markers.

Here, the input consists of candidates that were classified by the system as DCs, and a set of heuristic rules that analyse the syntactic structure of the elements automatically classified as term or definition is used to perform the ranking. Firstly, the ranking process assigns a numeric value to each identified term and definition of the candidates. Secondly, it combines those numeric values to generate a global value for each candidate.

Some of the heuristic rules can be seen in the next table:

Table 3. Example of ranking rules

Term = 1	<t>quotation marks .* quotation marks</t>
Term = 3	<t>.* pronoun .*</t>
Def = 1	<d>.* *that* .*</d>
Def = 3	<d>demostrative pronoun</d>

From the table above we can observe different rules that assign different values to the structure of terms and definitions. Value 1 means the best result, while 3 means the worst; candidates that do not follow any of the rules are assigned the value 2 by default. In the case of term's structures, the value 1 is assigned to those structures that are present between quotation marks, while a value of 3 is assigned to those candidates where the term structure consists of a pronoun, which could indicate a possible anaphoric reference. In the case of definition's rules, the value 1 is assigned to those structures where a relative clause is introduced after the pronoun *que* (that), which can be a prototypical structure in analytical definitions, while a value of 3 is given to the cases that consist only of a demonstrative pronoun. In the next table we illustrate some examples of each case:

Table 4. Example of ranking results

Term1	<t>«introns»</t>
Term3	<t>Este cloroplasto</t>
Definition1	<t>la mutación rutabaga</t> <dvp>es </dvp> <d>una mutación errónea que destruye a la adenilciclasa, interrumpiendo la síntesis del AMPc</d> .
Definition3	<d>Esto</d> <dvp>se conoce <nx> como</nx></dvp> <t>mutación</t>.

In the next sections we will describe our methodology for the system evaluation.

5. Evaluation

To develop the evaluation procedure we also used the IULA's technical corpus in Spanish. Taking into account that our system aims to identify DCs by searching for instances of definitional verbal patterns, we decided to set up a sub-corpus containing occurrences of the lemmas of the verbs from our grammar of verbal patterns. We searched for the first 250 occurrences of each verbal pattern (or all of the occurrences when they where less than 250), which produced a sub-corpus of 5809 sentences. Each one of those sentences was manually classified as DC or Non-Relevant, and was used as the input to the system to perform the evaluation.

We used precision & recall to evaluate the system performance. In this case, precision is the total number of

DCs automatically extracted, over the total number of candidates the system automatically identified as DCs, while recall is the number of DCs automatically extracted, over the total number of DCs presented in the evaluation sub-corpus.

The precision & recall results can be found in the next table:

Table 5. Precision & Recall results

P	R
0.53	0.79

It can be seen that almost the 80% of the total number of DCs was automatically extracted, while less than the 50% percent of the candidates was identified as noise, i.e., contexts that the system considers to be DCs but where manually tagged as Non-Relevant. In the case of recall, the system did not identify any candidates that were manually considered to be DCs.

In order to obtain a more specific scenario of the system's performance, we decided to apply an evaluation procedure for each kind of verbal patterns. For this purpose, we only considered those contexts containing one definitional verbal pattern. In this case, the sub-corpus consists of 4799 occurrences and the results are shown in the following table:

Table 6. Precision & Recall of definitional verbal patterns

Type	P	R
Analytical	0.58	0.83
Extensional	0.48	0.77
Functional	0.45	0.83
Synonimical	0.76	0.85

In general terms, it can be seen that the best results were obtained for synonimical patterns, while the lower values were obtained for the recall of the extensional patterns, and the precision for the functional patterns. These may be due to the fact that extensional patterns include verbs that can be used in a wider range of sentences and not only to introduce definitional information. Synonimical patterns, on the other hand, include verbs such as *conocer* (to known), *denominar* (to denominate), *llamar* (to call) and *nombrar* (to call) which, in conjunction with the particle *también* (also) seems to be more reliable for the recovering of definitional information. Analytical patterns show that some of the verbal forms can introduce a wider range of sentences that are considered to be noise. The same situation applies for the functional patterns.

6. Conclusions

We have presented a process of developing a definitional knowledge extraction system. This system aims at the simplification of the terminological practice related to the search for definitions of terms in specialised texts.

The methodology we have presented includes the searching for definitional patterns, the filtering of non-relevant contexts and the identification of DCs constituent elements: terms, definitions, and pragmatic patterns.

Up to now we have only worked with definitional verbs but we know that there is still further work to be done, which includes:

1. To explore other types of definitional patterns (mainly typographical patterns and reformulation markers) that are capable of recovering definitional contexts.
2. To improve the rules for the filtering process of non-relevant contexts, as well as to improve the algorithm for the automatic identification of constituent elements.
3. To improve the ranking algorithm.

Acknowledgments

This research was made possible by the financial support of the Consejo Nacional de Ciencia y Tecnología, Mexico and DGAPA-UNAM. The authors wish to thank the reviewers for its comments and suggestions, which helped to improve this paper.

7. References

[1] J. Pearson. Terms in Context. John Benjamin's, Amsterdam, 1998.

[2] I. Meyer. "Extracting Knowledge-rich Contexts for Terminography". In Recent Advances in Computational Terminology. D. Bourigault, C. Jacquemin and M.C. L'Homme (eds.).. John Benjamin's, Amsterdam, 2001. 278-302.

[3] J. Klavans and S. Muresan. "Evaluation of the DEFINDER System for Fully Automatic Glossary Construction". In Proceedings of the American Medical Informatics Association Symposium. ACM Press, New York, 2001. 252-262

[4] A. Storrer and S Wellinghoff. "Automated Detection and Annotation of Term Definitions in German Text Corpora". In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06). Genève, 2006. 2373-2376.

[5] A. Pinto and D. Oliveira. Extracção de Definições no Corpógrafo [on line]. University of Porto, 2004. http://www.linguateca.pt/documentos/OliveiraPintoOut2004.pdf

[6] R. Alarcón, C. Bach and G. Sierra. "Extracción de contextos definitorios en corpus especializados: Hacia una elaboración

de una herramienta de ayuda terminográfica". Revista Española de Lingüística. Madrid, 2008. 247-278.

[7] V. Malaisé, P. Zweigenbaum and B. Bachimont. "Mining Definitional contexts to Help Structuring Differential Ontologies". Terminology 11 (1), 2005. 21-53.

[8] S. Walter and M. Pinkal. "Automatic Extraction of Definitions from German Court Decisions". In Proceedings of the Workshop on Information Extraction Beyond the Document. 21st International Conference on Computational Linguistics (COLING'2006). Sydney, 2006. 20–28.

[9] P. Monachesi. "The LT4eL Project: Overview" [on line]. University of Utrecht. 2007.

www.lt4el.eu/content/files/ws_prague/lt4el-prague.pdf

[10] H. Saggion. "Identifying Definitions in Text Collections for Question Answering". In *Proceedings of the 4th International Conference on Language Resources and Evaluation LREC2004*. Lisboan, 2004. 1927-1930.

[11] R. Navigli and P. Velardi. "GlossExtractor: A Web Application to Automatically Create a Domain Glossary". In Lecture Notes in Computer Science 4733, 2007. 339-349

[12] F. Leu, and C. Ko. "An Automated Term Definition Extraction using the Web Corpus in Chinese Language". In Proceedings of the Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'07), 2007. 435-440.

Enriching a lexicographic tool with domain definitions: Problems and solutions

María A. Barrios
Universidad Complutense de Madrid
Paraninfo Ciudad Universitaria s/n
28008-Madrid. Spain
auxiba@filol.ucm.es

Guadalupe Aguado de Cea
Ontology Engineering Group – UPM
Avda. Montepríncipe s/n
28660-Boadilla del Monte. Spain
lupe@fi.upm.es

José Ángel Ramos
Ontology Engineering Group – UPM
Avda. Montepríncipe s/n
28660-Boadilla del Monte. Spain
jarg@fi.upm.es

Abstract

Enriching linguistic resources with domain information has been considered one important target in natural language applications. However, automatic definition extraction of this domain information from specialized resources has revealed certain methodological problems in definition construction. This paper presents some problems encountered in automatic definition extraction that are mainly related to inconsistencies in definitions, different granularity of definitions and embedded definitions. To face these problems some Meaning-Text Theory tools have been used: (a) semantic labels as a solution for inferring knowledge, (b) lexical functions as a way of providing coherence to definitions and (c) the actancial structure as a tool for developing consistent and complete definitions. Our goal is to describe the problems and to show the solutions proposed.

Keywords

Definition extraction, ontology building, linguistic resource enrichment, Meaning Text Theory.

1. Introduction

Reusing and enriching existing resources are nowadays two key issues both in academy and in the business world. In several scientific disciplines such as ontology development, computational linguistics, web semantic, ontologies and computational terminology the interest has been focused on many different aspects ranging from reusing lexicons, thesauri to create ontologies to extracting semantic relations from domain corpora or enriching definitions from specialized texts. One of the current drifts tends to build ontologies extracting definitions from different sources. However, building new resources with linguistic information extracted from different domain sources has revealed a difficult task as quite often the domain sources can be useful for a certain task but may show certain inconsistencies for others. In this paper, we present the problems encountered when trying to reuse three domain resources for two different purposes: (a) to build an ontology and (b) to populate a general linguistic resource, a database, with specific information from domain documents. With the aim of developing a consistent linguistic resource for further use in natural language applications, we focus on achieving consistent definitions of domain terms. Accordingly, we resort to the Meaning-Text Theory (MTT) principles [16] to propose some systematic solutions in order to avoid the inconsistency problems when building a terminological resource that can later be used in ontology development. Thus, we have mainly focused on three fundamental aspects: (a) semantic labels as a solution for inferring knowledge, (b) lexical functions as a way of providing coherence to definitions and (c) the actancial structure as a tool for developing consistent and complete definitions.

The rest of the paper is organized as follows: In section 2 we provide the scenario in which we have based our research and the tools used. Section 3 focuses on definition extraction and the pitfalls faced in the process. Section 4 presents a short review on definition typology. The MTT tools used and the database, BADELE 3000, are described in section 5. The problems encountered and the solutions proposed are presented in section 6. Finally, some conclusions are outlined in section 7.

2. Background

The domain resources used in this project summarized in this section (for more details, see Gómez-Pérez *et al* [7]) relate to geographic and geospatial information. All geographic information (GI) resources contain data about real entities and how to represent them in a map. So, each entity corresponds to an instance of a geographic phenomenon (*feature*). Indeed, the most important concept for GI is the *feature* since the Open GeoSpatial Consortium (OGC) [19] has declared that a geographic feature is the starting point for modelling geospatial information. In other words, a *feature*, which is the basic unit of GI, is an abstraction of a real world phenomenon associated with a location relative to the Earth, about which data are collected, maintained and disseminated [11]. Features can

Workshop On Definition Extraction 2009 - Borovets, Bulgaria, pages 14–20,

include representations of a wide range of phenomena that can be located in time and space such as buildings, towns and villages or a geometric network, a geo-referenced image, pixel or thematic layer.

For modelling this domain we have decided to use an ontology. To achieve this target, we have used three domain resources provided by the National Geographic Institute of Spain (IGN-E): the Concise Gazetteer (NC) -scale 1:1,000,000-, the Numerical Cartographic Database (BCN25) -scale 1:25,000-, and the National Topographic Database (BTN25) -scale 1:25,000-.

The Concise Gazetteer is a basic corpus of standardized toponyms created by the Spanish Geographical Names Commission. The first version has 3667 toponyms. This gazetteer complies with the United Nations Conference Recommendations on Geographic Names Normalization. The Concise Gazetteer has been created by the Spanish Geographical Names Commission. For further details, refer to Nomenclátor Geográfico Conciso de España [18].

The BCN25 presents an abstraction of reality, represented in one or more sets of geographic data, as a defined classification of phenomena. It defines the feature type, its operations, attributes, and associations represented in geographic data. For more information on this document see Rodriguez [21].

The BTN25 is the latest IGN-E catalogue and intends to be a sort of BCN25 reorganization, following a structure similar to frames. The instance information is the same as in BCN25, but the phenomena classification and its attributes are completely different.

These resources have one characteristic in common: each resource has a domain dictionary with phenomena. In the first case, NC phenomena, there is a txt file with 22 definitions. In the second case, BCN25 phenomena, an Excel file contains 366 definitions developed after the catalogue. Finally, there is a PDF document with "Capture rules for GI to be included in BTN25" (a first version), which describes its phenomena with 292 definitions (the document is not complete). In all cases, definitions were formulated by specialists on geography to facilitate the classification of the real entities in order to be included in the instance set of each resource.

All definitions are grouped by labels, as illustrated in Table 1 with four examples. These definitions have been used to build the ontology, as explained in section 3.

Table 1. INDUSTRIAL INSTALLATION (source document)

Nouns	Definitions
Corral (corral)	***Construcción*** *creada para cobijarse los pastores o para recoger el ganado* (**Construction** created for shepherds or cattle shelter)
Granja (farm)	***Hacienda*** *de campo que consta de establos, huerta y casa habitable* (**Ranch** with stables, an orchard and a house)
Piscifactoría (fish farm)	***Instalación*** *en la que se crían diversas especies de peces y mariscos con fines comerciales* (**Installation** where fish or seafood are bred for commercial purposes)
Palomar (pigeon loft)	***Edificio*** *donde se recogen y crían palomas* (**Building** where pigeons live and are bred)

3. Definition extraction

Definition extraction, as used in this paper, is the process of extracting the definition for a term from different resources. In our case these definitions have not been taken from corpora using machine learning techniques, as in many natural language processing applications [3], but from other domain resources with explicit definitions for these terms, their term variants or other semantically equivalent terms. However, some problems have appeared in this definition extraction process that showed certain inconsistencies and loss of information.

The definition extraction process followed to build and enrich a domain ontology is as follows: (1) the application we have developed retrieves the term from "Capture rules for GI to be included in BTN25"; (2) it extracts its definition from the same document; (3) it searches for the term in the auxiliary domain dictionaries; and (4) it extracts the corresponding definitions to add them to the corresponding classes. All these actions are executed automatically. Fig. 1 shows the overall workflow of information.

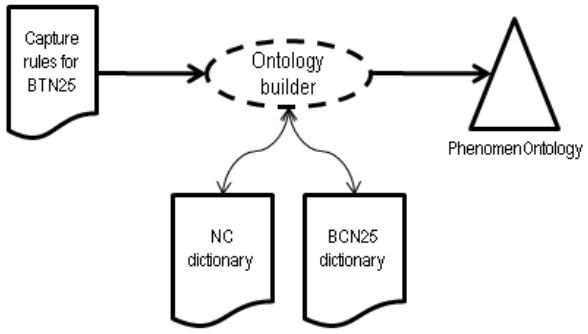


Fig. 1. Ontology building with definition extraction

As a result of this process, we obtained an ontology (called PhenomenOntology 3.5) which included 108 terms extracted from the documents mentioned and later transformed in 108 classes belonging to three groups: (a) classes without definitions; (b) classes with one definition; (c) classes with more than one definition. However, the

retrieval ratio of definitions extracted from the auxiliary dictionaries was very low, although they belonged to the same domain. In fact, only 4 definitions were found in the NC dictionary (although it contains 22 definitions, which means that 18 definitions were lost in the process) and 33 definitions were found in the BCN25 dictionary (it contains 366 definitions, which means that 333 definitions were also lost in the definition retrieval process).

The origin of this low ratio mainly lies on the abundance of terminological variants and semantically equivalent terms. For example, when trying to retrieve definitions for '*río*' (river) in the ontology, the system cannot recognize definitions of term variants such as '*río 1ª categoría*' (river 1st category) and '*río 2ª categoría*' (river 2nd category), and consequently it does not retrieve any of these definitions. Moreover, semantically equivalent terms are not retrieved when incorporating definitions in the ontology, as the system cannot recognize the similarity of the definitions of '*río*' (river) and '*corriente fluvial*' (flowing current).

Therefore, the problem is not only the loss of certain definitions in the extraction process but also the overlapping of some of them with different granularity which led to inconsistencies. For example, '*río*' (river) was retrieved with two definitions: *recorrido de una corriente de agua natural y de caudal más o menos constante, que recoge los aportes de una cuenca fluvial* (taken from the original document BTN25: "stream of natural water, with more or less constant flow, which collects water from other water courses") and *curso natural de agua* (taken from the NC dictionary: "waterstream").

Although these terminographic resources have been originally compiled by different experts, they show many lexico-semantic divergences that hinder the automatic definition extraction process. Quite often specific domain lexicographic resources are generally built to share information within a project team and attention is not usually paid to terminological principles when defining new terminology.

In other words, when building ontologies, automatic extraction of classes implies the annotation of these classes with definitions which are also automatically extracted. The final result of the definition extraction process reveals some problems that we have tried to tackle as explained in the next sections. Nevertheless, ontology building problems are out of the scope of this paper, though they have served as test bed for our research on principles for definition writing.

4. Definition typology

According to the traditional aristotelic genus-species definition, a definition should describe the concept and its relations to other concepts in the concept system. This type of definition is traditionally called formal definition, or intensional definition [8, 9]. That is to say, it reflects the superordinate concept to which the designation belongs and its delimiting characteristics. However, there are also other ways of designating concepts, extensional, ostensive, lexical, precising, and stipulative definitions [8] as well as ontological definitions [4]. For a more exhaustive revision on definitions see [13, 12]. Although these definitions can be useful for certain purposes depending on the user's needs and the approach adopted, they do not conform to a certain defining formulation and hinder any possibilities of formalizing the knowledge expressed in definitions in order to be used for natural language applications, such as knowledge extraction, ontology enrichment, to mention just a few. For this reason, we claim that some recommendations regarding terminological definitions should be considered when preparing domain resources. As [9, 10] stipulates the selection of an appropriate superordinate is crucial for the intelligibility of the defining statement. In Pearson's words [20] "the superordinate or closest generic concept should preferably be one step up in the hierarchy from the term being defined". Moreover, the same superordinate should be used for all terms that belong to the same class.

5. MTT lexicographic tools and BADELE.3000

In order to get more accurate systematic definitions, we decided to use the MTT tools. We considered two possible ways, (a) applying these tools directly to the ontology; (b) using them to enrich a general purpose lexicographic resource which could be later reused in other applications, for instance, for mapping the PhenomenOntology. At this point, we studied the advantages and disadvantages of the database BADELE.3000 [1, 2] that had been developed according to some MTT lexicographic tools.

BADELE.3000 is a database that contains the 3,000 most frequently used Spanish nouns. The information of each noun includes the definition and the combinatorial possibilities, among other linguistic information. A systematic process for the design of the database was followed; consequently the lexical data are well structured and separated from the applications that might use them. This way, the features of the data model and the subsequent database make them useful for different purposes, such as word sense disambiguation, machine translation and text generation.

As a result, the database contains a minimum of information useful for any type of ontology (because the general vocabulary includes some basic terms transversal to any specific domain) and more than 20,000 combinations. Besides, this resource allows us to infer knowledge potentially useful in real applications.

However, BADELE.3000 is a general-purpose resource with a low utility in commercial exploitations as it does not contain crucial information for real applications. The medium, long-term objective is to enrich this generic

linguistic resource by formalizing definitions which can help infer conceptual knowledge

Thus, our aim is twofold: To solve the problems of definition extraction and to add domain knowledge to a general purpose linguistic resource. The process followed is presented in Fig. 2.

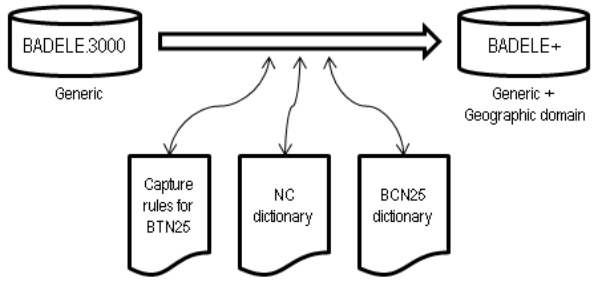


Fig. 2. Definition extraction and systematic lexicalization during BADELE upgrade

As for the lexicographic tools applied to BADELE.3000, we have resorted to three concepts proposed by the Meaning-Text Theory (MTT).

The first one is the *lexical function* (LF) [17: 39-40]: a LF associates a given lexical expression L (such as *sound*), which is the argument or keyword of F, with a set of lexical expressions –the value of F (such as loud, strong, heavy, deafening, etc). – expressing a specific meaning associated with F (for instance, 'intense' for the examples just mentioned which correspond to the LF known as **Magn**).

The second concept is the semantic label: a *semantic label* is the equivalent to the genus in traditional definitions by genus and differentia. For instance, *whale* could be defined as a 'sea mammal that breathes air through a hole at the top of its head and is hunted for meat and for other purposes, as a source of other materials'. The first part of this definition, 'sea mammal', the genus, is known in MTT approach as semantic label; the second part of this definition, the differentia, can be attached to some LFs.

The third concept is the actant [14, 15] and its derivate, the actantial structure. Actants correspond to beings or things that participate in the process expressed by a predicate: MTT approach considers that there is a sort of argument structure in all kinds of predicative words, which means that not only do the verbs have actants but also the adjectives, adverbs and the predicative nouns. The actantial structure reflects the syntactic expression of the actants, as shown in the example of *fleuve* (river) of Dicouèbe, in Table 2:

Table 2. *Fleuve* (river) Dicouèbe Actantial Structure

Nouns	Actantial Structure
Fleuve	[QUI COMMENCE AU lieu X, PASSE PAR LES lieux Z ET SE TERMINE DANS L'étendue d'eau Y]
River	[WHICH STARTS AT THE X place, FLOWS THROUGH THE Z places AND FINISHES AT THE Y area]

Among these three concepts, LFs have proved to be a specially helpful tool for lexicographic works such as the French dictionary *DECFC*[1], the French database *Dicouèbe*[2](developed in Montreal by Polguère and Mel'cuk) and the Spanish database *DiCE*[3] (developed in La Coruña by Alonso Ramos). Fontenelle [5] has also created (semiautomatically) a database but its originality derives from the fact that he takes as source bilingual dictionaries enriched with lexical-semantic information based on LFs. According to Frawley [6] the methodology followed by these resources is ideally suited to the compilation of specialized dictionaries.

6. Problems and solutions

In section 3 two problems have been pointed out when describing the definition extraction process. The low ratio of retrieved definitions can be solved by using linguistic resources (such as domain lexicons, WordNet, etc.) during the label search. So, term variants and semantically equivalent terms could be found and their definitions would be retrieved. The total number of definitions retrieved would increase. However, these definitions would show the same inconsistencies derived from the different granularity and specificity compared to existent ontology definitions. That is, the main problem in the whole process is the linguistic realization of definitions.

Thus, we have mainly focused on three subsidiary problems derived from the above mentioned problem and proposed some solutions according to MTT: (a) semantic labels as a solution for inferring knowledge, (b) lexical functions as a way of providing coherence to definitions and (c) the actancial structure as a tool for developing consistent and complete definitions.

6.1 Definitions and semantic labels

6.1.1 Problem: inconsistencies on the first part of definitions

The first problem that the technical definitions extracted from the knowledge resources used show is the inconsistencies between the name of the label of a group of terms (such as INDUSTRIAL INSTALLATION) and the first part of the definition, i.e. the superordinate of every single term under this label (such as construction, ranch, installation, building), because it differs from one to

[1] Information about the four volumes of this dictionary can be accessed at http://www.olst.umontreal.ca/decfr.html

[2] http://olst.ling.umontreal.ca/dicouebe/

[3] http://www.dicesp.com/

another, as Table 1 shows. The following question could be raised, why is a farm defined as a 'ranch', a corral as a 'construction', a fish farm as an 'installation' and a pigeon loft as a 'building'?

It is clear that the first part of every definition is used in an intuitive way as a quasi-synonym of the genus of the remaining definitions of the group. But in our view it is a false quasi-synonym. As a matter of fact, native Spanish speakers do not use ranch, building, container or installation as synonyms. This raises a second question, why all these words share the label but not the genus of the definition?

6.1.2 Solution: Semantic labels

The last question leads us to propose the use of semantic labels as envisaged in the MTT approach mentioned in section 5. A semantic label would correspond to the genus that matches the superordinate of the definition. Consequently, we propose the use of semantic labels as superordinates in the first part of the definition as a possible solution to avoid inconsistencies. In the examples in Table 2, we have used INDUSTRIAL INSTALLATION as a semantic label of the entire group, so all the definitions begin with the same superordinate. Table 3 shows our proposal.

Table 3. Our proposal for INDUSTRIAL INSTALLATION

Nouns	Definitions
Corral (corral)	***Instalación industrial*** *creada para cobijarse los pastores o para recoger el ganado* (**Industrial installation** created for shepherds or cattle shelter)
Granja (farm)	***Instalación industrial*** *que consta de establos, huerta y casa habitable* (**Industrial installation** with stables, an orchard and a house)
Piscifactoría (fish farm)	***Instalación industrial*** *en la que se crían diversas especies de peces y mariscos con fines comerciales* (**Industrial installation** where fish or seafood are bred for commercial purposes)
Palomar (pigeon loft)	***Instalación industrial*** *en la que se recogen y crían palomas* (**Industrial installation** where pigeons live and are bred)

6.2 Definitions and lexical functions

6.2.1 Problem: embedded definitions

Sometimes simple terms (nouns) or complex terms that share semantic features are defined differently. This inconsistency can be really subtle, as the example in Table 4 shows, based on the definitions of *bancal* (slope) and *ladera abancalada* (terrace slope).

Table 4. *Bancal* and *ladera abancalada* source definitions

Nouns	Definitions
Bancal (terrace)	*Rellano de tierra formado natural o artificialmente que frecuentemente se aprovecha para el cultivo* (Natural or artificial shelf that is frequently used for cultivation)
Ladera abancalada (terrace slope)	*Terreno pendiente con rellanos de tierra, naturales o artificiales, que se aprovecha para algún cultivo* (Natural or artificial terrace that is used for some kind of cultivation)

The two terms share all the semantic features, in other words, the basic characteristics. That would justify why the two definitions are almost equal. However, focusing on the object of the definitions, we find one definition is embedded in the other because a terrace slope is a set of slopes.

6.2.2 Solution: lexical functions

LFs are a powerful tool in order to give coherence to the definitions. Actually, the LF **Mult** could be quite useful in this and other similar cases. This LF expresses the sense 'set of X', where X is an argument that is usually filled by nouns, such as *grape*, or *flower*, as shown in (1):

(1) **Mult**(grape) = bunch of
Mult(flower) = bouquet of, bunch of

This LF can correspond to some lexical units that are not related syntagmatically (as examples above) but paradigmatically (in these cases, the value of the LF is preceded by the symbol //). Consequently, the final version of the entry of *bancal* in our database contains this LF, as shown in (2):

(2) **Mult**(bancal) = //ladera abancalada

The sense **Mult** is usually present at the beginning of definitions. For instance, the first sense of *bunch* is defined in the Oxford Dictionary as a number of things growing together, and the second one as a group of people. If we use the LF **Mult** in order to construct the definition, we should use *set of (bancales)* as the first part of *ladera abancalada*. Our proposal is shown in Table 5.

Table 5. *Bancal* and *ladera abancalada*: our proposal

Nouns	Definitions
Bancal (terrace)	*Rellano de tierra formado natural o artificialmente que frecuentemente se aprovecha para el cultivo* (Natural or artificial shelf that is frequently used for cultivation)

Ladera abancalada (terrace slope)	*Conjunto de bancales en terreno en pendiente* (Set of terraces on a slope)

6.3 Definitions and the actantial structure

6.3.1 Problem: different granularity in definitions

We have found definitions with different granularity in the domain resources used. This difference can derive from the fact that one definition is more explicit than another; or rather, it sometimes implies different entries in each document, such as *bus station* (present at BTN.25 document) and *depot station* (present at BCN.25 document), where *depot* is a hypernym of *bus*, as shown in Table 6.

Table 6. *Bus/depot station* definitions

Nouns	Definitions
Estación de autobuses (bus station) BTN.25	*Lugar donde hacen parada los autobuses para el trasiego de pasajeros y mercancía*s Place where buses stop for picking up and dropping off passengers and goods or freight
Estación de transportes (depot station) BCN.25	*Edificio en el que están las oficinas y dependencias de las diferentes empresas encargadas de conducir personas y cosas de un lugar a otro. También alberga el sitio donde habitualmente hacen paradas los vehículos* Building or place where different transport companies that pick up and drop off passengers as well as goods or freight have their offices. It also refers to the place where buses usually have conventional stops

In the second case, we have to decide if the definition should include the sense of 'offices of the enterprises', as appears in the second one, or not.

6.3.2 Solution: the actantial structure

The actantial structure is a helpful tool when writing definitions. Actually, if we regard the actantial structure of "bus station", in Table 7, we can see that each of the three actants is attached to some of the expressions, as shown in Table 8.

Table 7. *Bus station* actantial structure

Actantial structure	Bus Station X where the bus Y picks up the passengers Z

Table 8. Bus station actants and Spanish expressions

Actant	**Spanish expressions attached**
X (*place)*	*Estación de autobuses Méndez Álvaro* (Méndez Álvaro Bus station)
Y (*bus)*	*El autobús llega a la estación a las dos* (the bus arrives at the station at 2.00 o'clock)
Z (*passenger*)	*Juan coge el autobús de las dos* (John takes the bus at 2.00 o'clock)

As the complete sense of *bus station* is expressed by the three actants included in Table 8, we rule out the senses 'offices and locals of the enterprises'; then we add the semantic label ('place') and propose a definition quite close to the first one in Table 6, in which the actantial structure is contained, as shown in Table 9.

Table 9. Our proposal: *Bus station* definition

Nouns	Definitions
Estación de autobuses (bus station) BTN.25	*Local en el que paran los autobuses para la subida y bajada de pasajeros y mercancías* (Place where the buses stop for picking up and dropping off passengers and goods …)

7. Conclusions and Future work

MTT has shown the potential advantages of using a systematic approach for defining terms as it builds on the relations established among the relevant information included in definitions and it allows for some sort of semantic network formed with all the elements present in the definitions. In the process of definition extraction from the domain resources used two problems appeared: semantic inconsistency between different definitions for a concept (term), and very low efficiency of automatic definition search in auxiliary dictionaries. These problems have been described and some solutions have been proposed. Thus, we can conclude that MTT tools are very powerful in order to define or redefine terms. Semantic labels have proved to be consistent as superordinates; LFs are useful when choosing the essential sense of some definitions; and, finally, the actantial structure helps to complete other incomplete definitions.

As future work, our proposal would aim at developing an extraction methodology that could be documented in order to set the steps for automatic extraction. Thus, the manual process above mentioned could be described in detail as the problematic cases are identified and solved so as to identify all the possible activities than can be automatized. To sum up, the final objective is to build a framework which supports definition extraction as automatically as possible. This framework will help experts

in definition extraction and systematic lexicalization while adding domain knowledge to a generic lexicographic resource.

8. Acknowledgements

This work has been partially funded by the National Project "GeoBuddies: Anotación semántica colaborativa con dispositivos móviles en el Camino de Santiago" (TSI 2007-65677 C02) and the European Project "NeOn" (FP6-027595).

9. References

[1] Barrios Rodríguez, MA; Bernardos, MS. "*BaDELE.3000*: An implementation of the lexical inheritance principle". In Gerdes *et al*, (eds.) Meaning-Text Theory 2007. Proceedings of the 3nd International Conference on Meaning-Text Theory. Wiener Slawistischer Almanach. Sonderband, 69. 2007. Pages: 97-106.

[2] Bernardos, MS; Barrios, MA. "Data model for a lexical resource based on lexical functions". *Research in Computing Science*, vol. 27. 2008.

[3] Borigault, D. Jacquemin, C. & J'Homme, MC (eds.) Recent Advances in Computational Terminology, Amsterdam: John Benjamins, 2001.

[4] Cabré MT. *La terminología*. Barcelona: Empuréis. 1992

[5] Fontenelle T. "Using a Bilingual Dictionary to create Semantic Networks". Practical lexicography: a reader. Oxford. Oxford University Press. 2008. Pages: 169-190.

[6] Frawley W. "Lexicography and the Philosophy of Science" Dictionaries, 3:18-27. 1980/1981.

[7] Gómez-Pérez A, Ramos JA, Rodríguez-Pascual AF, Vilches-Blázquez LM. 'The IGN-E case: Integrating through a hidden ontology', The 13th International Symposium on Spatial Data Handling (SDH 2008), June 23rd - 25th, 2008. Montpellier, France. 2008. Pages: 417-435.

[8] ISO/DIS 704. Terminology work — Principles and Methods. 2008.

[9] ISO 1087-1. Terminology work. Vocabulary: Theory & Application. 2000.

[10] ISO 1087-2. Terminology work. Computer Applications. 2000.

[11] ISO 19110. Geographic Information – Methodology for feature cataloguing. 2005.

[12] Malaisé, V. Zweigenbaum, P. & Bachimont, B. 2005. "Mining defining contexts to help structuring differential ontologies". *Terminology*, Vol 11-1. 21-54.

[13] Martin, R. 1990. "La definition 'naturelle'." In Chaurand, J. & Mazières, F. (eds.) *La definition*. 86-95. Paris: Larousse.

[14] Mel'čuk I. "Actants in semantics and syntax I: Actants in semantics". *Linguistics*, 42:1, 2004a. Pages: 1-66.

[15] Mel'čuk I. "Actants in semantics and syntax II: Actants in syntax". *Linguistics*, 42:2, 2004b. Pages: 247-291.

[16] Mel'čuk, I and Polguère, A. 1987. "A formal lexicon in Meaning-Text Theory. Or how to do lexica with words". Computational linguistics. Nº 13, vol.3, 4, July-December, 1987. Pages: 261-275.

[17] Mel'čuk I. and Wanner, L. "Lexical functions and lexical inheritance for emotion lexemes in German". In Wanner, L. (ed.), *Lexical functions in lexicography and natural language processing*. Amsterdam/ Philadelphia. John Benjamin. 1996. Pages: 209-278.

[18] Nomenclátor Geográfico Conciso de España (versión 1.0). "Presentación y Especificaciones". Instituto Geográfico Nacional. Octubre 2006. http://www.idee.es/ApliVisio/Nomenclator/NGCE.pdf (ICC2005). A Coruña, Spain. 2006.

[19] OGC. OpenGIS Reference Model, Version 0.1.2, OGC Inc. Wayland, MA, USA. 2003.

[20] Pearson J. Terms in Contexts. Amsterdam/ Philadelphia: John Benjamins. 1998.

[21] Rodríguez Pascual AF, García Asensio L. "A fully integrated information system to manage cartographic and geographic data at a 1:25,000 scale". XXII International Cartographic Conference (ICC2005). A Coruña, Spain. 2005.

Extraction of Author's Definitions Using Indexed Reference Identification

Marc Bertin, Iana Atanassova and Jean-Pierre Descles
Paris-Sorbonne University
Maison de la Recherche
28 rue Serpente
75006 Paris
{marc.bertin / iana.atanassova / jean-pierre.descles}@paris-sorbonne.fr

Abstract

In this paper we present the implementation of definition extraction from multilingual corpora of scientific articles. We establish relations between the definitions and authors by using indexed references in the text. Our method is based on a linguistic ontology designed for this purpose. We propose two evaluations of the annotations.

Keywords

Semantic annotation, definition extraction, indexed references

1 Introduction

The use of definitions plays an important role in a number of scientific disciplines. The complexity of some domains raises the necessity to develop tools for the automatic annotation of information relevant to definitions. The growth in scientific literature production leads us to propose new tools for text navigation and quick access to the textual information.

In this paper we explore a new way to extract definitions from scientific text corpora by establishing a relation between the usage of a definition and a cited author.

In section 2, we describe the elaboration of a linguistic ontology, based on the analysis of multilingual corpora. Then, the identification of indexed references is used to establish the relations between authors.

In section 3, we explain our implementation. The goal of our system is to provide to the user the possibility to clarify a notion and its usage in a given context from a terminological or conceptual viewpoint. This means that we need to maintain the link between the extracted definitions and their contexts, in order to provide access to the argumentation in the text. The user can thus visualise the context in which the term in question has been defined. Section 4 shows the results produced by the application. In section 5 we discuss the problem of the evaluation of the semantic annotations and propose two types of evaluations: one by the precision/recall measures and another by the Cohen's weighted Kappa coefficient.

Finally, we conclude by a discussion of the perspectives for the utilisation of this tool.

2 Methodology

We propose a method for the identification of definitions and also for the identification of relations between authors. This approach allows us to associate a definition to an author and to establish a link with other texts that could interest the user. The system allows a fully automated text processing, which comprises several stages.

2.1 Protocol

Our protocol is as follows: first we carry out the identification of the sentences containing indexed references, by using regular expressions. Then, we annotate the definitions in the sentences identified in the previous stage. Finally, we extract the definitions and create indexes for the information retrieval. The results are stored in a database. Different types of visualizations and information retrieval are provided by our web-based interface.

2.2 Multilingual Corpora

We have constructed multilingual corpora, in order to create our linguistic resources organized in a linguistic ontology. The corpora comprise mainly scientific texts and articles available online. The French corpus consists of texts from several scientific reviews (Intellectica, ALSIC, TALN, IRISA) and six PhD theses from the domains of Linguistics and Computer science. The articles in English corpus are from Nature, Journal of Cell Science, Biophysical Journal, Proceedings of the National Academy of Sciences, The Journal of Cell Biology, and others.

Corpus	Texts	Sentences
French	205	119410
English	116	38378
Total	321	158788

Table 1: *Corpora*

In table 1 we present the sizes of the corpora. In order to ensure compatibility with the tools of segmentation and annotation, the corpora have been converted into text files. The sentence counts are obtained after

Workshop On Definition Extraction 2009 - Borovets, Bulgaria, pages 21–25,

the segmentation, which will be detailed later in the section 3.2.2.

From a legal point of view, texts can be cited freely, even if under copyright[1], provided that the following three criteria are respected. Firstly, citations must be short: our interface provides output in the form of text segments corresponding to sentences. Secondly, the purpose of extraction must be infromative, such as in the case of information retrieval. Finally, the source must be mentioned.

Moreover, we establish a relation between the definition and the cited document or author through the bibliography. This stage is important for the creation of an author network.

2.3 Definition Ontology

This section describes our linguistic approach and the construction of an ontology for the annotation of definitions. The method we present is based on enunciative discourse considerations and a corpus analysis, through which we construct an ontology by abduction.

2.3.1 Linguistic Analysis

We can examine a definition sentence by studying the relation between the *definiendum*, what is to be defined, and the *definiens*, what defines it. This linguistic study of our corpus has led us to a better understanding of the distinction between a *definition* and a *definatory characteristic*, which has been taken in consideration for the construction of our linguistic resources. We define a definatory characteristic as a sentence that gives only some essential properties of the defined object. We have distinguished three categories of definatory characteristics: *identification, determined categorization* and *pseudo-definition*. We have also considered two sub-categories of the definition: general definitions and axiomatic definitions. The full ontology that we have created contains some further sub-categorizations that are presented on figure 1.

The categorization in this linguistic ontology is based on an analysis of the types of relations. Here we will describe briefly the differences between some of the categories that we have retained.

Firstly, it must be noted that in definition sentences, apart from the relation between the definiendum and the definiens, there exists a second relation, which is between this first relation and the agent who established the definition. The presence of this agent is not always manifested in discourse and sometimes there is no actual trace. In the case when the agent is present in the text, we can speak of a *contextualised definition*, because it is often marked in the context by a deictic, which is limited to a domain or to a period in time, or else introduced by a passive construction or 'on' in French.

Secondly, we have *axiomatic definitions*, which are utterances expressing a primary truth.

Finally, there are cases where the author uses a reported definition. In these cases the enunciator can choose whether to attest the definition or not, in order to use it in the elaboration of a demonstration, or to introduce a new notion. This type of definitions takes part in the text evolution by means of modalities and we speak of *committed definition*.

The objective that we have fixed is to extract definition sentences, in which the definition is explicitly attributed to an author or another work, cited in the text. We will also call them *signed definitions*, which correspond to the category of Reported Definitions in our ontology.

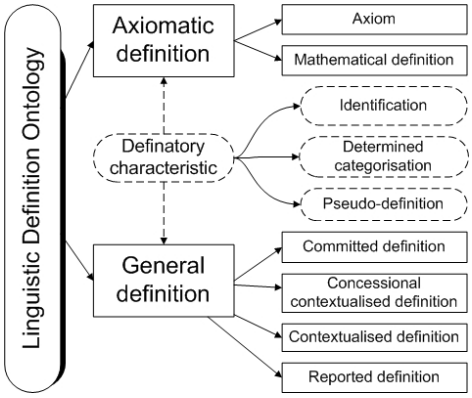


Fig. 1: *Linguistic ontology of the definition*

2.4 Indexed Reference Identification

The method that we propose is based on the indexed references in the text which point to the bibliography as define in [1]. More precisely, the indexed references allow us, in the case when we identify a definition in the research scope determined by the segmentation, to link this definition to the author cited in the text. The theoretical framework as well as the experimental procedures for the indexed reference identification are described below.

2.4.1 Bibliographic International Standards

We have considered several norms for bibliographic references, namely the norms ISO-690 and ISO 690-2, which are the international standards from the International Organization for Standardization, as well as the French norms AFNOR NF Z 44-005 and AFNOR NF Z 44-005-2.

In practice the norms are not rigorously applied by authors of scientific texts. For this reason, a method based only on the norms described above is not sufficient to carry out the text processing on a large scale. That is why, although the identification of the indexed references may seem trivial at first glance, a large number of morphological and syntactic variations must be taken into account. To illustrate this complexity, here is a list of forms that we have extracted from our corpus: *(Hoc, 1990a), (Thom, 1970), (Dingwall et al., 1995; Hartmann and Görlich, 1995), [24], Pickett-Heaps et al. (1990), (like other authors e.g. Raven, 1983), (Cwuc and SPRAGUE 1989), (18, 53, 56).*

[1] cf. CPI art L. 122-5

2.4.2 Finite State Automata

Although the identification of indexed references has been approached by Citeseer, we have developed our own module. In fact, at the beginning of our work such modules were not available[2]. The specificity of our module is its capacity to identify also the author names which can appear in the forms. The classification that we use has been published in [2].

We identify automatically the indexed references by the use of Finite State Automata (FSA). For this we have to take into consideration the norms established on the one hand by the practices proper to authors and on the other hand by the different domains. That is why in order to create robust FSA, different corrections had to be made to take into consideration the different customs in writing indexed references. The annotation platform we have chosen takes as input rules based on lists of regular expressions. Therefore, for the implementation of this methodology, we have converted the FSA into regular expressions.

2.4.3 Identification of Known Named Entities

The identification of an indexed reference can become difficult because of the presence of named entities in the reference. The named entities are the more complex part of the indexed references and introduce considerable complications in the FSA due the various name morphologies in different languages.

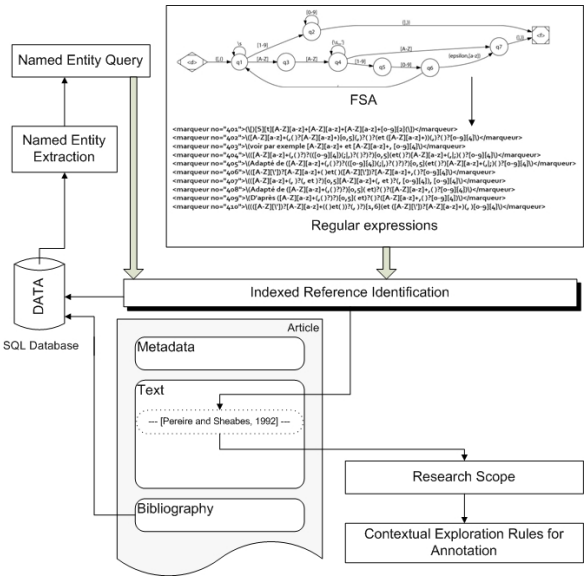


Fig. 2: *Indexed reference identification with Named Entity extraction*

Figure 2 describes the implementation of a solution which improves the system performance, through the utilization of author names, that have been already identified by the system, as part of the regular expressions. In fact, by using some data already existing in the bibliographic databases, we can generate certain forms and limit the noise in the more complex forms. Moreover, this approach permits some extensions of the method: we can consider new sentences for the signed definition extraction through matching not only indexed references, but also author names in the text that can be cited without bibliographic links.

3 Semantic Annotation

3.1 Annotation Tools

Definition identification is traditionally based on pattern matching, as for example in [11]. These approaches are used for the development of platforms such as TerminoWeb[3] of the National Research Council Canada.

Different approaches are possible for the semantic annotation. Among the tools that we have considered we can cite the GATE[4] platform [12] based on machine learning algorithms, generally used with JAPE [4], and the work of Xerox Concept-matching, based on XIP [10], a morphosyntactic analyser

In our work we have used the Excom platform [6], which implements the Contextual Exploration method [5]. This is a decision-making procedure, presented in the form of a set of rules and linguistic markers that trigger the application of the rules. They are applied to the segments containing indicators. The indicators are linguistic units that carry the semantic meaning of the categories for annotation. After the initial identification of the indicators in the text, the rules carry out the localisation of complementary linguistic clues which are co-present in the context of the indicators. After the verification of the presence or absence of the linguistic clues, the rules attribute a semantic annotation to the segment.

In our approach we consider as a working hypothesis the fact that in a scientific article the information related to signed definition can be found in the textual space close to an indexed reference, and more specifically in the same sentence. Our aim is to limit as much as possible the noise in the annotations, to be able to obtain foolproof matching between authors and definitions.

As we need to be able to disambiguate the linguistic forms according to the context, in order to limit the noise as much as possible and to deal with polysemy, we have chosen the Contextual Exploration framework as more adapted to our approach. For this reason, we have used the Excom annotation system[5].

3.2 System Overview

Here we describe in detail the main stages in the text processing, that we have divided into a four-stage process. The overall system pipeline is presented on figure 3.

[2] The source code of the Citeseer module has been recently published on http://sourceforge.net/projects/citeseerx/. The identification module is based on a Perl module from CPAN, which parses documents using regular expressions.

[3] http://termino.iit.nrc.ca/
[4] http://gate.ac.uk
[5] http://www.excom.fr

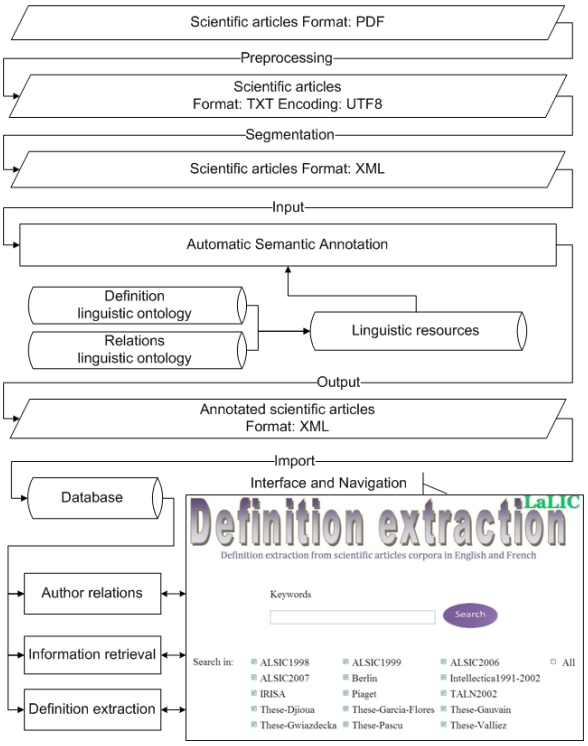


Fig. 3: *Stages in the automatic processing*

3.2.1 Preprocessing

The initial corpora being in PDF format, in the preprocessing stage the files are converted into text format. This is necessary because the next stages in our processing, namely the segmentation and annotation, need full text access to the corpora. The converted files are in the UTF8 encoding which permits the processing of different natural languages.

3.2.2 Segmentation

In the second stage, we segment the corpora into paragraphs and sentences, in order to prepare the input for the annotation module. The quality of the segmentation is important for the overall system performance, as the segmentation provides the text elements to be annotated. The segmentation is carried out by the SegaTex module [7] which we have chosen for the reliability of its results and its capacity to process texts in both English and French. This module takes as input text files and returns the segmented files in the Docbook XML format, with paragraph and sentence elements. This format is compatible with the annotation module.

3.2.3 Semantic Annotation

We will therefore analyze the discourse forms which are to be found in the text space close to the indexed reference. The semantic annotation is carried out by the Excom module which takes as input the segmented files as well as the linguistic resources that we have constructed. According to our protocol, we use two types of linguistic resources: regular expressions for the identification of the indexed references and contextual exploration rules for the annotation of definitions related to the ontology presented above. In the output, the identified indexed references are present as new elements and the annotations of the definitions are added as attributes to the relevant XML sentence elements in the corpora.

3.2.4 Interface and Navigation

We have developed a web-based graphical user interface, using the technology Apache/PHP/MySQL. The annotated corpora are imported into a database designed for this purpose, which contains the annotations and the text segments, as well as meta-data related to the files. The Definition Extraction Interface (DEI) permits the visualization of the information in the database, and different other functionalities that we will describe here.

The most important functionality of the DEI is the information retrieval among the annotated sentences. In the initial screen, shown on figure 3, the user can formulate a query by using keywords[6] and eventually restricting the search to a specific set of corpora. The results are presented in the form of a list of sentences, together with the annotations and links to the initial texts.

3.3 Results

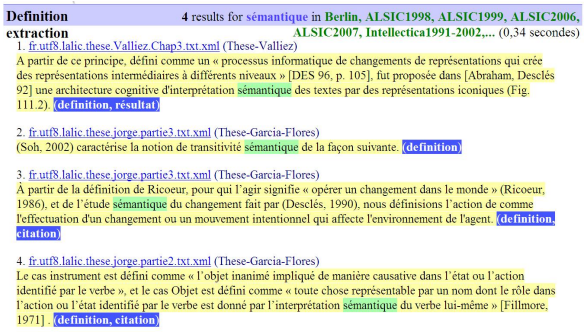


Fig. 4: *Search results*

Figure 4 presents the results from the French corpus for the keyword *"sémantique"*. The following excerpts were extracted from the English corpus:

1. Another homolog to RCCI has been identified in S. cerevisiae, called either SRMl (Cwuc and SPRAGUE 1989) or PRP20 (AEBI et al. 1990; FLEISCHMANN et al. 1991).

2. Silica polymerization occurs within an organelle called the silica deposition vesicle, bounded by a membrane called the silicalemma (18, 53, 56).

We can see that the first and the second examples are general reported definitions.

[6] Boolean expressions (AND, OR, NOT) with parentheses and quotation marks in queries are also implemented.

4 Evaluation and Discussion

4.1 Precision and Recall Measures

The first evaluation consists in measuring the accuracy of the retained indexed references, which have been identified automatically by the regular expressions. We have used the precision/recall measures [9] which determine the capacity of the system to correctly identify textual segments containing indexed references. Table 2 presents the number and the percentage of the sentences containing indexed references in each corpus. We can see that around 5% of the sentences have been extracted.

Corpus	Sentences	Annotated Sentences	Percentage
French	119410	5976	5,00 %
English	38378	1743	4,54 %
Total	157788	7719	4,89 %

Table 2: *Annotated sentences*

We have carried out the evaluation on a set of 500 sentences extracted randomly from our corpora. In table 3 we present the results obtained by this evaluation.

Recall	Precision	F-measure
0,911%	0,989%	0,9483

Table 3: *Evaluation of the Indexed References*

We consider that these results are satisfactory. It must be noted that there is very little noise which means that almost all of the identified indexed references are valid. On the other hand, the value of the recall is also very high. The several percents of indexed references not identified by the system are due to the various orthography rules for the names in different languages, as well as the presence of commentaries in the indexed reference itself.

4.2 Cohen's Weighted Kappa

The problem we have to consider is how to evaluate the semantic annotation which is by definition qualitative in nature. The test Kappa (K) proposed by Cohen[3] and developed by [8] provides a method to measure numerically the agreement between two or more observers or methods in the case when the judgments are qualitative in nature. We have adopted this method for the second stage of our evaluation.

		Judge A		
	Reponses	Correct	Incorrect	Total
Judge B	Correct	33	5	38
	Incorrect	3	9	12
	Total	36	14	50

Table 4: *Evaluation Results*

In order to carry out the test, we have constituted a base of annotated text segments and these segments have been evaluated independently by two human judges. The judges had to classify the segments into two categories: correct and incorrect. We have used a set of 50 sentences for this evaluation. Table 4 presents the results. For the Cohen's Kappa we obtain: $\kappa = 0,6515$, and therefore we have a subtantial agreement, according to the interpretation in [8].

5 Conclusion and Future Work

We note that according to the evaluation the system gives satisfactory results, which validates the linguistic resources and the definition ontology in our approach. Throughout the process of annotation and exploitation of the results we maintain the links between the extracted sentences and the original texts which makes possible the visualization of the context of each definition. The evaluations confirm the relevance of this application. However, we are not yet able to predict the result on a larger scale and on corpora in other domains. In the future we will extend this approach to the processing of bigger corpora in English and in French as well as other natural languages.

References

[1] M. Bertin. Categorizations and annotations of citation in research evaluation. In *FLAIRS 2008, Coconut Grove, Florida*, Coconut Grove, Florida, May 2008.

[2] M. Bertin, I. Atanassova, and J.-P. Desclés. Automatic analysis of author judgment in scientific articles based on semantic annotation. In *22nd International Florida Artificial Intelligence, Research Society Conference*, Sanibel Island, Florida, 19-21 mai 2009.

[3] J. Cohen. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 20:27–46, 1960.

[4] H. Cunningham, D. Maynard, and V. Tablan. Jape–a java annotation patterns engine. *Advances in Text Processing, TIPSTER Program Phase II*, pages 185–189, 2000.

[5] J.-P. Desclés. Contextual exploration processing for discourse automatic annotations of texts. *FLAIRS 2006, Florida. Invited Speaker*, 2006.

[6] B. Djioua, F. J. Garcia, A. Blais, J.-P. Desclés, G. Guibert, A. Jackiewicz, F. L. Priol, L. Nait-Baha, and B. Sauzay. Excom: an automatic annotation engine for semantic information. *FLAIRS 2006, Florida*, pages 285–290, 2006.

[7] M. Ghassan. La segmentation de textes par exploration contextuelle automatique, présentation du module segatex. *ISLsp, Inscription Spatiale du Langage : structure et processus IRIT, Université Paul Sabatier, Toulouse*, 2002.

[8] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.

[9] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1979.

[10] A. Sandor, A. Kaplan, and G. Rondeau. Discourse and citation analysis with concept-matching. *International Symposium : Discourse and document (ISDD), Caen, France*, 2006.

[11] G. Sierra, R. Alarcon, C. Aguilar, and C. Bach. Definitional verbal patterns for semantic relation extraction. *Terminology*, 14(1):74–98, 2008.

[12] V. Tablan, C. Ursu, K. Bontcheva, H. Cunningham, D. Maynard, O. Hamza, T. McEnery, P. Baker, and M. Leisher. A unicode-based environment for creation and use of language resources. In *Proceedings of 3rd Language Resources and Evaluation Conferenc*, 2002.

Evolutionary Algorithms for Definition Extraction

Claudia Borg
Dept. of I.C.S.
University of Malta
claudia.borg@um.edu.mt

Mike Rosner
Dept. of I.C.S.
University of Malta
mike.rosner@um.edu.mt

Gordon Pace
Dept. of Computer Science
University of Malta
gordon.pace@um.edu.mt

Abstract

Books and other text-based learning material contain implicit information which can aid the learner but which usually can only be accessed through a semantic analysis of the text. Definitions of new concepts appearing in the text are one such instance. If extracted and presented to the learner in form of a glossary, they can provide an excellent reference for the study of the main text. One way of extracting definitions is by reading through the text and annotating definitions manually — a tedious and boring job. In this paper, we explore the use of machine learning to extract definitions from non-technical texts, reducing human expert input to a minimum. We report on experiments we have conducted on the use of genetic programming to learn the typical linguistic forms of definitions and a genetic algorithm to learn the relative importance of these forms. Results are very positive, showing the feasibility of exploring further the use of these techniques in definition extraction. The genetic program is able to learn similar rules derived by a human linguistic expert, and the genetic algorithm is able to rank candidate definitions in an order of confidence.

Keywords

Definition Extraction, Genetic Algorithms, Genetic Programming.

1 Introduction

Definitions provide the meaning of terms, giving information which could be useful in several scenarios. In an eLearning context, definitions could be used by a student to assimilate knowledge, and if collected in a glossary, they enable the student to rapidly refer to definitions of keywords and the context in which they can be found. Unfortunately, identifying definitions manually in a large text is a long and tedious job, and should ideally be automated. In texts with strong structuring (stylistic or otherwise), such as technical or medical texts, the automatic identification of definitions is possible through the use of the structure and possibly cue words. For instance, in most mathematical textbooks, definitions are explicitly marked in the text, and usually follow a regular form. In less structured texts, such as programming tutorials, identifying the sentences which are definitions can be much more challenging, since they are typically expressed in a linguistically freer form. In such cases, humans have to comb through the whole text manually to tag definitional sentences.

One way of automating definition extraction is to consult human linguistic experts to identify linguistic forms definitions conform to, usually using either lexical patterns or through specific keywords or cue-phrases contained in the sentence. Once such rules are identified, automatic tools can be applied to find the sentences matching one or more of these forms. This approach has been shown to work with varying results. Technical texts fare better than non-technical ones, where results are usually not of a satisfactory level. Two issues which limit the success of these results are (i) the relative importance of the different linguistic forms is difficult to assess by human experts, and is thus usually ignored; and (ii) coming up with effective linguistic forms which tread the fine line between accepting most of the actual definitions, but not accepting non-definitions, requires time and expertise and can be extremely difficult. Since there typically is a numeric imbalance between definitions and non-definitions in a text, having a slightly over-liberal rule can result in tens or hundreds of wrong positives (non-definitions proposed as definitions), which is clearly undesirable. In the approach we propose, we give a degree of importance (weight) to each linguistic form. Through this technique, one could go further than simple human-engineered linguistic forms — by being able to rank the sentences by how probable the system thinks they are actual definitions. The more a sentence matches against the more important forms, the higher the degree of confidence in its classification as a definition.

In this paper, we explore the use of machine learning techniques, in particular evolutionary algorithms, to enable the learning of sentence classifiers, separating definitions from non-definitions. We have used two separate algorithms for two distinct tasks:

- *Relative importance of linguistic forms:* Given a number of predetermined linguistic forms which definitions may (or usually) conform to, we have used a genetic algorithm to learn their relative importance. Through this technique we enable a more fine-grained filter to select definitions, taking into account multiple rules, but at the same time assigning them different weights before performing the final judgement. We thus benefit from having a ranking mechanism which would indicate a level of confidence in the classification of the definitions. In a semi-automated scenario, it would make the system more usable since a human expert would be presented with the best re-

Workshop On Definition Extraction 2009 - Borovets, Bulgaria, pages 26–32,

sults first, and results are grouped by 'quality' of the definition.

- *Learning the linguistic forms:* The previous technique assumes that we start off with linguistic forms which are able to match definitions — a task which would typically require human expert input. We incorporated genetic programming techniques to learn such forms automatically by generating different rules in the classification task. Within such a setup it is possible to explore new linguistic structures and test their worthiness automatically against the training data. Rule which are found to be useful in classifying definitions are kept and improved upon to evolve to a better solution.

These two separate techniques are then combined to provide us with a fully automated definition extraction system by first identifying a number of linguistic forms through the use of genetic programming, and then using the genetic algorithm to assign to each rule a degree of importance. The resulting features and their associated weights can then be used by a definition extraction tool which will not only extract candidate definitional sentences, but also rank them according to a level of confidence. The results achieved when combining these two techniques are very promising, and encourage further investigation of these techniques in the task of automatic definition extraction.

In section 2 we give a short overview of definition extraction and the setup of our experiments. In section 3 we describe the results of the genetic algorithm experiment, while in section 4 we present the genetic programming experiments and results achieved. In section 5 we discuss how these two components can be merged into one complete definition extractor, and compare the results to other related work in this area in section 6. We then conclude and discuss future directions in section 7.

2 Definition Extraction

Rule-based approaches to definition extraction tend to use a combination of linguistic information and cue phrases to identify definitions. For instance, in [12, 14] the corpora used are technical texts, where definitions are more likely to be well-structured, and thus easier to identify definitions. Other work attempts definition extraction from eLearning texts [17, 13] and the Internet [6]. Non-technical texts tend to contain definitions which are ambiguous, uncertain or incomplete compared to technical texts.

In our work, we focus on definition extraction from non-technical eLearning English texts in the field of ICT. The corpus consists of a collection of learning objects gathered as part of the LT4eL project [11] which were collected from several tutors in different formats, and standardised in XML format. It is generally recognised that part-of-speech information, which can be extracted automatically from natural language texts is crucial to enable effective discrimination, and the corpus is thus annotated with linguistic information, using the Stanford part-of-speech tagger [15].

The corpus was manually annotated with definitions, to be used as a training set for the definition extraction task. Manually crafted grammars were created in the project to extract definitions, however the results were not satisfactory [1]. From observation it was noted that the structure of definitions does not always follow a regular *genus et differentia* model and different styles of writing and definitions pose a major challenge for the identification of definitions. The solution adopted was to categorise the definitions into different classes, and engineer definition recognisers for each of the classes separately. This reduces the complexity, by attempting to identify a grammar focusing for each type of definition. The types of definitions observed in the LT4eL texts were classified as follows:

1. Is-a: Definitions containing the verb 'to be' as a connector. E.g.: 'A joystick is a small lever used mostly in computer games.'

2. Verb: Definitions containing other verbs as connectors such as 'means', 'is defined' or 'is referred to as'. E.g.: 'the ability to copy any text fragment and to move it as a solid object anywhere within a text, or to another text, usually referred to as cut-and-paste.'

3. Punctuation: Definitions containing punctuation features separating the term being defined and the definition itself. E.g.: 'hardware (the term applied to computers and all the connecting devices like scanners, telephones, and satellites that are tools for information processing and communicating across the globe).'

Three further categories have been identified and used in the LT4eL project, but were not considered for our experiments due to the difficulty of applying machine learning in those instances.

3 Definition Extracting using Genetic Algorithms

Definition extraction is usually based on a set of rules which would have been crafted by a human linguistic expert. The rules would usually contain the discriminating features between definitions and non-definitions, and can be generic (in the form of `Noun Phrase` · `verb to be` · `Noun Phrase`) or very specific part-of-speech sequences. Experts usually identify different rules, some of which may be overlapping (that is, a sentence may match more than one rule). Combining such rules can enable more effective definition extraction. At its simplest level, one can adopt the policy which gives preference to sentences which match more of the rules: a sentence matching five rules would be preferred than a sentence matching two rules. However not all rules are equally effective in identifying definitions. Ideally one would want to assign a weight to each rule indicating its relative importance. The setting of these weights can be performed using machine learning techniques.

3.1 Genetic Algorithms

A Genetic Algorithm (GA) [5, 4] is a search technique which emulates natural evolution, attempting to search for an optimal solution to a problem by mimicking natural selection. By simulating a population of individuals (potential solutions) represented as strings, GAs try to evolve better solutions by selecting the best performing individuals (through the use of a *fitness function*), allowing only the best individuals to survive into the next generation through reproduction. This is done using two operations called *crossover* and *mutation*. Crossover takes two individuals (parents), splits them at a random point, and switches them over, thus creating two new individuals (children, offspring). Mutation takes a single individual and modifies it, usually in a random manner. The fitness function measures the performance of each individual[1], which is used by the GA to decide which individuals should be selected for crossover and mutation, and which individuals should be eliminated from the population. This process mimics survival of the fittest, with the better performing individuals being given higher chances of reproduction than poorly performing ones, and thus their winning characteristics are passed on to future generations.

In our work, we have explored the use of a GA to learn the weights to a predetermined set of linguistic rules. These weights will represent the relative importance of the respective rule in its effectiveness at classifying definitions.

3.2 Combining Features

A feature is considered to be a test which, given a sentence s, returns a boolean value stating whether a particular structure, word or linguistic object is present in the sentence — essentially, characteristics that may be present in sentences. These could range from rendering information (bold, italic), to the presence of keywords, or part-of-speech sequences that could identify the linguistic structure of a definition. So, if we take the presence of a bold word to be a feature for definitional sentences, then a sentence containing a bold word is more likely to be a definition than a sentence which does not.

Given a vector of n basic features, $\overline{f} = \langle f_1, \ldots f_n \rangle$, and numeric constants, $\overline{\alpha} = \langle \alpha_1, \ldots \alpha_n \rangle$, one can define a compound feature combining them in a linear manner:

$$F_{\overline{\alpha}}^{\overline{f}}(s) = \sum_{i=1}^{n} \alpha_i \times f_i(s)$$

Given a sentence, a vector of features and their respective weights, we can thus calculate a numeric value of the sentence by combining the features accordingly. One would also have to identify a threshold value τ such that only sentences scoring higher than this value would be tagged as definitions i.e. s is tagged as a definition if and only if $F_{\overline{\alpha}}^{\overline{f}}(s) \geq \tau$.

[1] The fitness of an individual is the measure of how good this candidate solution is at solving the problem being tackled.

3.3 Learning Weights

We have used a GA to identify a good set of weights and the threshold value for a given set of features. Each individual in the population of the genetic algorithm is represented as the vector of numeric weights. Crossover between individuals simply consists of splitting the vector of the two parents at a random position, and joining the parts, thereby creating two new individuals for the next generation.

What the GA learns is determined by the fitness function, which, given an individual, returns a score of how 'good' the individual is. We have used a corpus of definitions and non-definitions to evaluate the performance of each individual. The fitness function takes an individual (vector of weights) and runs through the whole corpus using the combined feature function and calculates how many definitions are correctly classified, and how many are incorrectly tagged as non-definitions. Similarly, we compute the values for the non-definitional sentences. Through these figures we are then able to extract precision, recall and f-measure. We have run the GA using these different measures as the fitness function.

The choice of threshold is obviously crucial to the value returned by the fitness function. One option was to set the threshold to a fixed value for the whole population (say, at zero). However, it was noted that given an individual one can actually compute an optimal value for the threshold with respect to the corpus using an efficient (linear) algorithm. Another option was to include the threshold as part of the individual's chromosome. However, this would have serious implications on the effectiveness of the learning unless the crossover function is defined in a more careful manner, since during crossover one would mix-and-match weights of individuals with different thresholds. Combining two good individuals would typically result in a non-effective one in this manner. We opted not to explore this option.

Two experiments were run, one with a fixed threshold value of zero, and another using optimal (individual specific) thresholds, with the latter achieving far better results.

3.4 Experimental Results

The GA experiments focused on the 'is-a' category, where we had 111 definitions and 21,122 non-definitional sentences. Several experiments were carried out, using different techniques within the algorithm mechanics. The best selection algorithm was SUS with sigma scaling [10]. Here we present a summary of the best and most interesting results of this work.

During the set up of the GA, we used a simple set of ten features which were hand-coded and inputted into the GA for it to learn their relative importance. Following is the set of features used:

1. contains the verb "to be"
2. has sequence "IS A" ("to be" followed by a determiner)

3. has sequence "FW IS" (FW is a tag indicating a foreign word - in the example "The process of bringing up the operating system is called booting", booting is tagged as an FW.)
4. has possessive pronoun (I, we, you, they, my, your, it)
5. has punctuation mark in the middle of the sentence (such as a hyphen or colon)
6. has a marked term (keyword)
7. has rendering (italic, bold)
8. has a chunk marked as an organisation
9. has a chunk marked as a person
10. has a chunk marked as a location

These features were purposely simplistic when compared to the manually crafted rules in the LT4eL project for definition extraction. This enabled us to analyse the relative weights assigned and to be able to allow more focus on the algorithmic aspects of the GA. These features were used throughout all the experiments discussed in this section.

Table 1: *Results for best experiments*

Method	F-measure	Precision	Recall
Experiment 1	**0.57**	0.62	0.52
Experiment 1a	0.62	**0.70**	0.42
Experiment 1b	0.54	0.46	**0.56**
Experiment 2	**0.57**	0.64	0.50
Experiment 3	**0.54**	0.59	0.50

Table 1 presents the results achieved by the best performing runs, indicating the f-measure, precision and recall achieved by assigning the weights learnt to the set of features. The best runs achieved an f-measure of 57%, with the runner-up achieving 54%. Since we used f-measure as the basis of measuring the weights' effectiveness in classifying definitions, we were also able to influence f-measure to favour precision or recall according to the setting of the alpha value. Experiments 1a and 1b show the results for favouring precision and recall respectively.

Using a small set of simple features, the GA has managed to obtain positive results, especially when comparing to the manually crafted grammars in LT4eL. We have increased precision from 17% to 62%, whilst maintain recall over 50%. Further improvement would probably be achieved had we to include more rules from the manually crafted grammar as part of our set of features.

The possibility of influencing the learning of weights to favour precision or recall is considered a positive facility in this experiment, since the end use of the definition extraction tool could require different settings. In a fully automatic system, precision might be given more importance, whilst in a semi-automatic system, recall is more important since a human expert will verify the correctness of the candidate sentences.

```
feature         ::=  simplefeature
                 |   simplefeature & feature

simplefeature   ::=  lobj
                 |   emptystring
                 |   any
                 |   simplefeature ?
                 |   simplefeature *
                 |   simplefeature . simplefeature
                 |   simplefeature + simplefeature
```

Fig. 1: *Specification of the representation of individuals*

4 Feature Extraction using Genetic Programming

The main bottleneck of using the GA as discussed in the previous section is that the linguistic rules have to be identified by a human expert. From the LT4eL experience it was clear that linguistic experts were needed to identify complex rules which non-experts would not have identified. The rules identified by experts are typically expressed as complex grammars or regular expressions ranging over parts-of-speech. In this section we present another experimental setup we have used to explore the use of machine learning techniques for the automated identification of linguistic rules.

4.1 Linguistic Rules

Recall that linguistic rules are objects which given a sentence, return a boolean value, depending on whether or not the sentence matched the rule. One way of expressing such rules is through the use of regular expressions, e.g. `noun·is·a·noun`. These regular expressions would range over the grammar shown in figure 4.1.

Note that the basic elements of the regular expression are simple linguistic objects (with no structure). Note also that to enable more complex rules, we allow not only the usual regular expression operators (optional inclusion, repetition, catenation and choice), but also allow the conjunction of regular expressions at the top most level (thus controlling the computational complexity of matching the regular expression).

The framing of basic features as instances of this language of regular expressions, enables us to formulate the task of the learning algorithm as that of learning an instance of this language (of regular expressions) which is effective when used for definition extraction.

For the choice of linguistic objects, we chose to either use specific part-of-speech tags such as NN (noun, common, singular or mass) or to generalise these tags into one class and refer to them as nouns.

4.2 Genetic Programming

Genetic programs (GP) are another form of evolutionary algorithms introduced by [8] whose aim is that of

automatically learning instances of a language, typically computer programs, automatically. The algorithm is very similar to GAs in structure — it is a search optimisation technique, exploring different possible solutions to a problem. Similarly to a GA, this technique uses crossover and mutation to evolve new individuals, and a fitness function to test the strength of the individual. One of the main differences between the two techniques is that unlike GAs, a GP uses tree representation to represent the individual.

Several of the definition extraction tasks tend to use rules made of part-of-speech information, which is generally arrived to through linguistic expertise or through observation of definitional sentences and their linguistic structure. In the process of creating such rules it is usually not very clear as how to best tweak a rule for better performance. Thus, an experimental setup which would create rules automatically and test them upon an annotated corpus is desirable. When a rule created is able to match correctly a sentence, it is kept as a potentially good rule to use in a definition extraction tool. A GP is an ideal experiment for this task as it facilitates the process of rule discovery and tests their effectiveness through the evolutionary process.

Since the evolutionary process is based on matching sentences against the rules created, we have also used f-measure as the fitness metric to determine whether a rule (an individual) is a good possibility or not. Those rules which have a higher f-measure will be kept by the GP so as to explore similar possibilities.

4.3 Experimental Results

Experiments using the GP delved into the three categories identified in section 2, that is the is-a, verb and punctuation categories. For each category, several experiments were run, each of which resulting in different rules (albeit at times quite similar). Experiments also tested the inclusion of different linguistic objects by either focusing on specific POS tags, or by generalising the particular category, say to include all nouns. In the case of the verb category, some of the experiments focused on the POS tags, while others included certain words such as 'known', 'define' and similar words typically found in definitions in this category.

Table 2 shows a summary of the best results achieved in the different categories where the GP was applied. The experiments were run with different population size ranging from 200 to 1,000 individuals. Most of the experiments converged within 100 generations, and at times as early as 30 generations. The selection of the individuals to survive to the next generation used elitism (which copies the best individuals of the population into the next generation as is), and selecting the remaining individuals for crossover using the stochastic universal sampling algorithm [10]. Further details about the experimental setup can be found in [2].

The GP was able to learn at times rather simple rules such as `noun·is·a·noun`. The rules learnt for each category by the different experiments were usually similar in structure and content. However in certain runs the rules represented by the individuals gave better results. In the is-a category, the average f-measure obtained was around 25%, with one run managing to produce a slightly different rule achieving 28% f-measure. In the verb category it was noticed using part-of-speech categories was not sufficient, and that the use of keywords, such as 'know', 'define', and 'call', was necessary to achieve good results. In the punctuation category we observed that results were achieved easily primarily due to a smaller search space when compared to the other categories.

Table 2: *Summary of results*

Category	F-measure	Precision	Recall
Is-a	0.28	0.22	0.39
Verb	0.20	0.14	0.33
Punctuation	0.30	0.25	0.36

5 Combining the Experiments

The two experiments described above were so far isolated, each one with a particular purpose. The challenge towards which we worked is to have a fully automated definition extraction tool which is easily adaptable to different domains and which ranks candidate definitions according to some level of confidence. In this section we describe how these two separate experiments were combined together towards a fully automated definition extraction tool. In figure 3 we see the different phases of the definition extraction process. Phase one is the creation of an annotated training set and is not dealt with in this work. Given an annotated corpus with definitions, one can then move onto phase two where the GP is applied to learn useful simple features which can be used to distinguish definitions from non-definitions. In phase three the GA is then used to learn weights for the rules learnt by the GP. Using the rules and weights, one can incorporate all this in a definition classification tool in phase four. In this section we present the results achieved from combining phase two and three together.

For the purpose of combining the two phases, we used the best rules learnt by ten different GP experiments in the is-a category. These individuals were used by the GA to learn their respective weights. The set up is shown in figure 2 where the final result is a set of rules in the is-a category together with their allocated weights indicating the level of effectiveness each weight has. As shown in the previous section, the rules the GP learnt without the application of the weights resulted at best in f-measure being 28%. Once weights were learnt and applied to the definition extraction tool, this increased to 68% f-measure. This improvement shows that learning weights is useful to the classification task since it does matter which rule is actually carrying out the classification of sentences.

Further analysis show that the f-measure is resulting from a 100% precision and a 51% recall. This means that by combining the rules learnt and their associated weights, we succeeded in classifying just over half of the annotated definitions, without classifying any incorrect definitions. There are several factors behind these results:

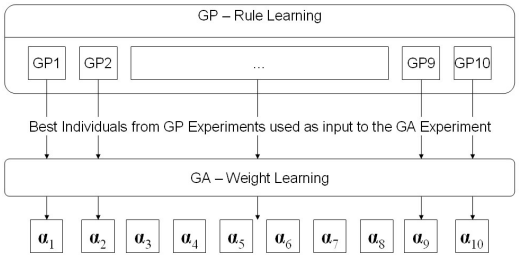


Fig. 2: *Combining the two experiments*

1. This experiment was carried out on only one corpus, so the rules learnt together with their respective weights, were specific to the corpus used. Achieving such a good result is only indicative that as in any machine learning process, the two algorithms were able to learn rules and weights specific to our corpus.

2. The recall of 51% represents definitions for which the genetic program did not learn rules for. Since these algorithms are searching for solutions in an automatic manner without expert feedback, it is the case that not all possible rules are explored. This can be tackled by including rules from more experiments or by having direct feedback from a linguistic expert (say, injection of good humanly crafted rules into the population).

Notwithstanding the conditions under which they were achieved, the results are very promising.

6 Discussion and Related Work

Although the results achieved so far are promising and encourage further investigation of these techniques, it is difficult to provide a fair and just comparison to other techniques. One of the main reasons is that an evaluation using an unseen corpus is required to have a more realistic view of the results achieved using these techniques. To our knowledge there is no other work in definition extraction using evolutionary algorithms to which our results can be directly compared to.

However, there are various attempts at definition extraction using different techniques. DEFINDER [12] is a rule-based system which extracts definitions from technical medical texts so that these can later be used in a dictionary. The rules are primarily based on cue-phrases such as "is called a", with the initial set of candidate sentences being filtered out through the use of POS rules and noun phrase chunking. They manage to obtain a precision of 87% and a recall of 74%. Definition extraction is also considered to extract the semantic relations present in definitions. In [9], they apply lexico-syntactic patterns in addition to cue phrases, focusing on hypernym and synonym relations in sentences. They obtain 66% precision and 36% recall.

Work carried out in [14], applies valency frames to capture definitional sentences achieving an average of 34% precision and 70% recall across the rules created. A German corpus consisting of legal decisions is used

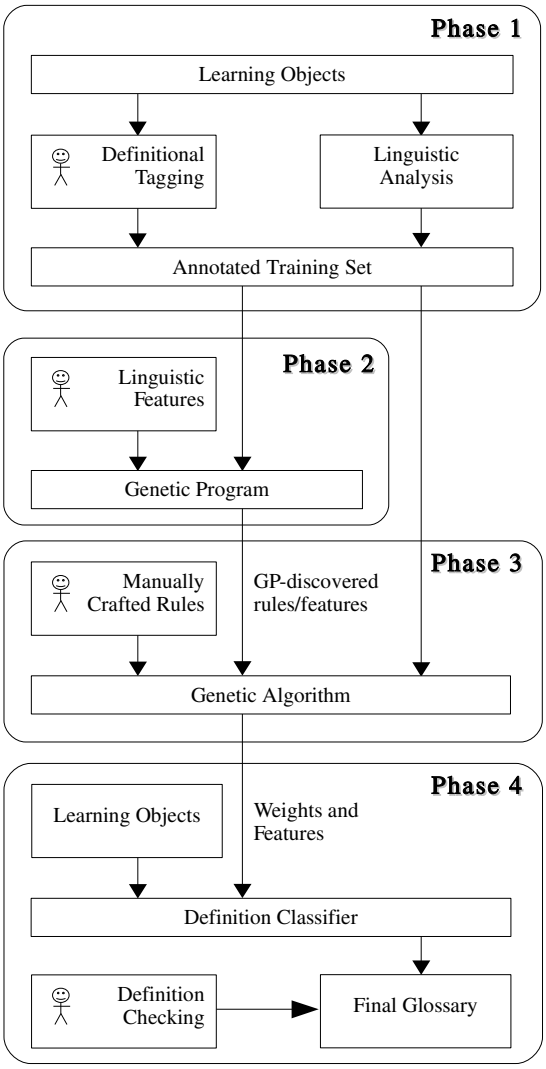


Fig. 3: *Phases of definition extraction*

in [16] to extract definitions. They analyse the structure of definitions in this domain, and observe that the German word `dann` can be used as a signal word indicating that a sentence is a definition. There is no equivalent term in English. The rules are crafted manually through observation, and achieve an average of 46% precision. When only the most effective rules are used, precision increases to over 70%, however recall is not discussed since the corpus is not annotated with definitions. Extraction of definitions from eLearning texts is attempted for the Slavic group of languages in [13], using noun phrase chunking and phrase structure as the potential identifying features in definitions. The best results are achieved for the Czech language with precision at 22% and recall at 46%.

Research in general seems to point out to the need of going beyond rule-based techniques, and trying out machine learning to improve definition extraction. Definitions extracted from the Dutch Wikipedia from

medical articles in [3] first use a rule-based approach using cue-phrases, but further improve their extraction process by using Naive Bayes, maximum entropy and SVNs. As part of their feature set they include sentence positioning, a feature which cannot be applied to other types of corpora. The best result is from applying maximum entropy, achieving 92% accuracy. Similar experiments by [17] on an eLearning corpus obtain 88% accuracy, with the difference in result being due to the type and structure of the corpus used. Similarly [7] obtain an accuracy of 85% using a Balanced Random Forest on an eLearning corpus. These techniques all share the similarity in having improved considerably the results of manually crafted grammars when applying machine learning techniques.

7 Future Directions

In this paper, we have presented a methodology for the use of evolutionary algorithms to create sentence discriminators for definition extraction. We have shown how GPs can be used to learn effective linguistic rules, which can then be combined together using weights learnt through the use of a GA. The overall system can, with very little human input, automatically identify definitions in non-technical texts in a very effective manner. Using our approach we have managed to learn rules similar to the manually crafted ones by the human expert in the LT4eL project, and further associate them with weights to identify the definitions in non-technical texts — all performed in an automated fashion. One of the major strong points of the approach is that the (expensive) learning phases is performed once, and the resulting definition discriminator is very efficient, making it viable to be included in other applications.

The final experiment of using both techniques for definition extraction gave surprising results, managing to identify only definitions, achieving a 100% precision, albeit having identified rules to capture only half of the definitional set of sentences. This result is certainly encouraging when considering that the process is fully automated.

There are various directions we plan to explore in the future. Our experiments would need to be evaluated further, experimenting with other corpora in different domains. For instance, medical texts contain several terms which a part-of-speech tagger might not recognise and would tag as 'foreign word'. Thus the rules learnt for our eLearning corpus might not necessarily apply for a medical corpus.

We also intend to evaluate the definition extraction tool over an unseen corpus. Such an evaluation might show that the rules learnt by the GP are not generic enough to cover unseen definitions, a result which is common in such machine learning techniques. It would be ideal to have some form of feedback loop from an expert to the learning algorithm to integrate new knowledge gained over unseen corpora.

We plan to explore and assess the use of weights to go beyond a crisp discriminator, and interpret the results as a fuzzy discriminator, associating a degree of confidence with each sentence, thus enabling us to rank definitions according to how sure the system is that it is a definition. This is crucial if the definitions discovered are to be vetted by a human operator.

Finally, we plan to extend the use of GP to learn rules in an iterative manner. After each iteration of the experiment, the sentences for which it learnt rules are removed from the training corpus, and the experiment repeated. In this way we would be reducing the search space, and forcing the GP to learn new rules. It might be the case that the GP does not learn certain rules as they would classify to many non-definitions to simply capture few definitions. However, by carrying out such an experiment we might be able to learn rules which cover the search space better, and at the same time identify those definitions for which it is difficult to define rules which provide acceptable results.

References

[1] C. Borg. Discovering grammar rules for Automatic Extraction of Definitions. In *Doctoral Consortium at the Eurolan Summer School 2007, Iasi, Romania.*, pages 61–68, 2007.

[2] C. Borg. Automatic Definition Extraction Using Evolutionary Algorithms. Master's thesis, University of Malta, 2009.

[3] I. Fahmi and G. Bouma. Learning to Identify Definitions using Syntactic Features. In *Workshop of Learning Structured Information in Natural Language Applications, EACL, Italy*, 2006.

[4] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA, 1989.

[5] J. H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, 1975.

[6] J. L. Klavans, S. Popper, and R. Passonneau. Tackling the internet glossary glut: Automatic extraction and evaluation of genus phrases. In *SIGIR'03 Workshop on Semantic Web*, 2003.

[7] Ł. Kobyliński and A. Przepiórkowski. Definition Extraction with Balanced Random Forests. In *proceedings of GoTAL*, 2008.

[8] J. R. Koza. *Genetic Programming: On the Programming of Computers by means of Natural Selection*. MIT Press, Cambridge, MA, 1992.

[9] V. Malaisé, P. Zweigenbaum, and B. Bachimont. Detecting semantic relations between terms in definitions. In *COLING CompuTerm 2004: 3rd International Workshop on Computational Terminology*, pages 55–62, 2004.

[10] M. Mitchell. *An Introduction to Genetic Algorithms*. MIT Press, 1998.

[11] P. Monachesi, L. Lemnitzer, and K. Simov. Language Technology for eLearning. In *First European Conference on Technology Enhanced Learning*, 2007.

[12] S. Muresan and J. L. Klavans. A method for automatically building and evaluating dictionary resources. In *Proceedings of the Language Resources and Evaluation Conference*, 2002.

[13] A. Przepiórkowski, Ł. Degórski, M. Spousta, K. Simov, P. Osenova, L. Lemnitzer, V. Kubon, and B. Wójtowicz. Towards the automatic extraction of definitions in Slavic. In *Proceedings of the BSNLP workshop at ACL*, 2007.

[14] A. Storrer and S. Wellinghoff. Automated detection and annotation of term definitions in german text corpora. In *Language Resources and Evaluation Conference*, 2006.

[15] K. Toutanova and C. D. Manning. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), Hong Kong*, 2000.

[16] S. Walter and M. Pinkal. Automatic Extraction of Definitions from German Court Decisions. In *Workshop on Information Extraction Beyond The Document*, pages 20–28, 2006.

[17] E. Westerhout and P. Monachesi. Extracting of Dutch Definitory Contexts for elearning purposes. In *CLIN*, 2007.

Language Independent System for Definition Extraction: First Results Using Learning Algorithms

Rosa Del Gaudio António Branco
University of Lisbon
Faculdade de Ciências,Departamento de Informática
NLX - Natural Language and Speech Group
Campo Grande, 1749-016 Lisbon, Portugal
rosa@di.fc.ul.pt antonio.branco@di.fc.ul.pt

Abstract

In this paper we report on the performance of different learning algorithms and different sampling technique applied to a definition extraction task, using data sets in different language. We compare our results with those obtained by handcrafted rules to extract definitions. When Definition Extraction is handled with machine learning algorithms, two different issues arise. On the one hand, in most cases the data set used to extract definitions is unbalanced, and this means that it is necessary to deal with this characteristic with specific techniques. On the other hand it is possible to use the same methods to extract definitions from documents in different corpus, making the classifier language independent.

Keywords

machine learning, imbalanced data set, language independent, definition extraction

1 Introduction

According to Aristotle, the formal structure of a definition should resemble an equation with the *definiendum* (what is to be defined) on the left hand side and the *definiens* (the part which is doing the defining) on the right hand side. The *definiens* should consist of two parts: the *genus* (the nearest superior concept) and the *differentiae specificae* (the distinguishing characteristics). In this way, definitions would adequately capture the concept to be defined.

In Hebenstreit [9], two more types of definition are pointed out. Firstly, the definition by enumeration of the concept species on the same level of abstraction (extensional definition), e.g. a chess piece is a king, a queen, a bishop, a knight, a rook or a pawn. Secondly, the definition by enumeration of the parts of the concept (partitive definition), e.g. the solar system is made of the planets Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune and Pluto. Barnbrook [2] identifies 16 different types of definitions analysing dictionary entries. In spite of the richness of this classification, in automatic definition extraction application only the simplest type is taken in consideration, that is a sentence composed by a subject, a copular verb and a predicative phrase. In this paper a definition is a sentence containing an expression (the *definiendum*) and its definition (the *definiens*) connected by the verb "to be".

Two different approaches are possible when dealing with automatic definition extraction. The first one consists in building a system of rules, based on lexical and syntactic clues. The second one is to consider the task as a classification problem, where for each sentence in the corpus it is possible to assign the correct class. The problem of the first approach is that it is language dependent, and in case of a large use of lexical clues, the performance on different corpus get worst. In the case of classification approach one of the main issue to be dealt with is the sparseness of definitions in a corpus. It is a matter of fact that the number of definition bearing sentences is much lesser than the number of sentences that are not definitions. This configuration gives rise to an imbalanced data set, which may present different degrees of imbalance, depending on the corpus used. For corpus composed mostly by encyclopedic documents it is likely to get a balanced data set. For example [8] used a balanced corpus where the definition-bearing sentences represent 59% of the whole corpus, while [24] using a corpus consisting of encyclopedic text and web documents reports that only 18% of the sentences were definitions.

In this work we deal with the problem of imbalanced data sets in definition extraction tasks in a language independent way. We show not only that sampling techniques can improve the performance of classifiers but also that this improvement is language independent. Other researches using learning algorithms relay strongly on lexical and syntactic components as features to describe the data set. These kinds of features are not only language dependent but also domain dependent, and as we want our classifier to be as general as possible we select the most basic features, that is n-grams of part of speech (POS). This makes the present approach viable for all those languages that are not equipped with rich lexical resources as learning data or in a situation where the domain is too specific to benefit from such resources, and moves away from previous works that use features such as words, word lemmas, position of the sentence in the document he document, etc. In this paper we apply the same techniques we applied to a Portuguese Corpus in a previous experiment to a corpus in Dutch and compare results. Our task handles several aspects that are common to

Workshop On Definition Extraction 2009 - Borovets, Bulgaria, pages 33–39,

different machine learning tasks in NLP application: small amounts of data, inherent ambiguity (definition detection is sometimes a matter of judgment), noisy data (human annotators make mistakes), imbalanced class distribution, this last aspect being the main issue addressed in this paper.

2 Related Work

As we said in the previous section there are two main approach to deal with automatic definition extraction, the rule based and the classification one. Regarding the first approach Hearst [11] proposed a method to identify a set of lexico-syntactic patterns to extract hyponym relations from large corpora and extend WordNet with them. This method was adopted by [19] to cover other types of relations.

DEFINDER [13] is considered a state of the art system. It combines simple cue-phrases and structural indicators introducing the definitions and the defined term. The corpus used to develop the rules consists of well-structured medical documents, where 60% of the definitions are introduced by a set of limited text markers. The nature of the corpus used can explain the high performance obtained by this system (87% precision and 75% recall).

Malaise and colleagues [16] focused their works on the extraction of definitory expressions containing hyperonym and synonym relations from French corpora. These authors used lexical-syntactic markers and patterns to detect at the same time definitions and relations. For the two different relations (hyponym and synonym), they obtained, respectively, 4% and 36% of recall, and 61% and 66% of precision. Turning more specifically to the Portuguese language. Pinto and Oliveira [20] present a study on the extraction of definitions with a corpus from a medical domain. They first extract the relevant terms and then extract definition for each term. An evaluation is carried out for each term; for each term recall and precision are very variable ranging between 0% and 100%.

In the last years machine learning techniques were combined with pattern recognition in order to improve the general results. In particular, [8] used a maximum entropy classifier to extract definition in order to distinguish actual definitions from other sentences. As attributes to classify definition sentences they used-such as n-gram and bag-of-words, sentence position, syntactic properties and named entity classes. The corpus used was composed by medical pages of Dutch Wikipedia, where they extracted sentences based on syntactic features. The data set were composed by 2,299 senteces of which 1,366 actual definitions. This gives an initial accuracy of 59%, that was improved with machine learning algorithms until 92.21%

In [6], it is presented a system to extract definition from off-line documents. They experimented with three different algorithms, namely NaïveBayse, Decision Tree and Support Vector Machine (SVM), obtaining the best score with SVM with a a F-measure of 0.83 with a balanced data set.

In [26] they combine syntactic patterns with a Naïve Bayes classification algorithm with the aim of extracting glossaries from tutorial documents in Dutch. They use several properties and several combination of them, obtaining an improvement of precision of 51.9% but a decline in the recall of 19.1% in comparison with a the syntactic pattern system developed previously by the authors using the same corpus.

Recently, some authors have started to look at this problem of imbalanced data set in the context of definition extraction. In particular, [21] down-sampled their corpus using different ratios (1:1, 1:5, 1:10) in order to seek for best results. The corpus they used presented an original ratio of non-definitions to definitions of about 19. Although they obtained some improvement in terms of F-measure, in particular with the ratio 1 to 5, they cannot improve results obtained with a rule based grammar previously developed using the same corpus. These authors also investigated the use of Balanced Random Forest algorithm in order to deal with this imbalance, succeeding in outperform the rule based grammar previously developed of 5 percentage points [14].

3 Corpora

All the two corpora used for experiments were collected in the context of the LT4eL project [1]. They were used to develop different tools, such a key-word extractor, a glossary candidate detector and an ontology, in order to support e-learning activities[1] in a multi-language context. The corpora are encoded with a common XML format. The DTD of this format is conforming to a DTD derived from the XCESAna DTD, a standard for linguistically annotated corpora [18]. Definition-bearing sentences were manually annotated. In each sentence, the term defined, the definition and the connection verb were annotated using a different XML tag.

The Dutch Corpus is composed by 26 tutorials with a size of about 350,000 tokens. The corpus was annotated part-of-speech information and morphosyntactic features with the Wotan tagger and with lemmatization information with the CGN lemmatizer (for more information about this corpus see [26].

The Portuguese Corpus is composed by 23 tutorials and scientific papers in the field of Information Technology and has a size of 274,000 tokens. It was then automatically annotated with morpho-syntactic information using the LX-Suite [23] a set of tools for the shallow processing of Portuguese with state of the art performance.

In order to prepare the data set for to be used in our experiments a simple grammar for each language was create that extracts all the sentences where the verb "to be" appears as the main verb. For Dutch we obtained a sub-corpus composed by 4,829, 120 of which are definitions, with a ratio of 39:1. For Portuguese we obtained a sub-corpus composed by 1,360 sentences, 121 of which are definitions, with a ratio of about 10:1.

Commonly used features are: bag-of-word, n-grams [17] (either of part-of-speech or of base forms), the position of the definition inside the document [12], the presence of determiners in the *definiens* and in the *definiendum* [8]. Other relevant properties can be the

[1] www.lt4el.eu

presence of named entities [8] or data from en external source such as encyclopedic data, wordnet, etc. [22].

Some features work well with a corpus but not so well in a different corpus, resulting in the impossibility to use the learner with different corpora. The use of the position of a definition-bearing sentence in [8] is an example of a feature that is corpus dependent. The same issue arise when lexical information is used as feature. In order to avoid such limitation we represented instances as n-grams of POS. From both the corpora the 100 most frequent n-grams were extracted and were used as features. Each sentence was represented as an array where cells record the number of occurrences of these n-grams. In this paper, for question of space, only results obtained with the best representation are showed, that is with bi-grams.

4 Machine Learning Algorithms

Five different algorithms were used: C4.5, Random Forest, Naïve Bayes, k-NN, SVM. The reason that motivated this choice is twofold: we want to cover different class of algorithms and we want to use algorithms representing the state of the art for definition extraction.

C4.5 and Random Forest are two decision tree algorithms. The first is a relatively simple algorithm that splits the data into smaller subsets using the information gain in order to chose the attribute for splitting the data. The second is a classifier consisting of a collection of decision trees. For each tree, it is selected a random sample of the data set (the remaining is used for error estimation) and for each node of the tree, the decision at that node is based on a restricted number of variables. Regarding C4.5, different configuration were tested: reduced-error pruning instead of C.4.5 pruning, pruned and unpruned option, and with or without Laplace smoothing. Regarding Random Forest, we experimented with different numbers of randomly chosen attributes.

Naïve Bayes is a simple probabilistic classifier that is very popular in natural language application. In spite of its simplicity, it permit to obtain results similar to the results obtained with more complex algorithms. Two different implementation were tested: one in which the numeric estimator precision values are chosen using a kernel estimator for numeric attributes and another using a normal distribution.

The k-NN algorithm is a type of instance-based learning, also called lazy learning because, differently from algorithms above, the training phase of the algorithm consists only in storing the feature vectors and class labels of the training samples and all computation is deferred until the classification phase. In this phase, it computes the distance between the target sample and n samples in the data set, assining the most frequent class. Two different K nearest neighbors classifiers were constructed, with k equal to 1 and to 3.

SVM is a classifier that tries to find an optimal hyperplane that correctly classifies data points as much as possible and separate the point of two classes as far as possible. In this experiment four different classifiers were implemented, using four different kernels, linear, polynominal, radial and sigmoid.

Weka workbench [27] was used to build all the learners.

5 Sampling Techniques

In many real-world classification applications, most of the examples are from one of the classes, while the minority class is the interesting one. As most of the learning algorithms are designed to maximize accuracy, the imbalance in the class distribution leads to a poor performance of these algorithms. The issue is therefore how to improve the classification of the minority class examples. A common solution is to sample the data, either randomly or intelligently, to obtain an altered class distribution.

Random over-sampling consists of random replication of minority class examples, while in random down-sampling majority class example are randomly discarded until the desired amount is reached. These two very simple methods are often criticized due to their drawbacks. Several authors pointed out that the problem with under-sampling is that this method can discard potentially useful data that could be important for the induction process. On the other hand, Random over-sampling can increase the likelihood of overfitting, since it makes exact copies of the minority class examples.

When speaking about negative and positive example in a dataset, it is important to have in mind that not all the examples have the same value. There are examples that are more prototypical than others and represent better the class to which they belong, others are too similar to be useful, and others are just noise.

It is possible to divide examples in four different classes:

- Noise examples - examples that are incorrectly classified
- Borderline examples - dangerous since a small amount of noise can make them fall on the wrong side of the decision border.
- Redundant examples - too similar to other examples to be useful.
- Safe examples - examples that fit perfectly the class to which they belong.

Building on these considerations, several methods were proposed in order to retain safe examples in the re-balanced data set. We present here two of such methods, namely the Condensed Nearest Neighbour Rule and Tomek Link algorithm.

Condensed Nearest Neighbor Rule [10] finds a consistent subset of examples in order to eliminate the examples from the majority class that are distant from the decision border, since these examples might be considered less relevant for learning. A subset $E' \subset E$ is consistent with E if using a 1-nearest neighbor,E' correctly classies the examples in E. First, it randomly draw one majority class example and all examples from the minority class and put these examples in E'. Next, it uses a 1-NN over the examples in E' to classify the

examples in E. Every misclassified example from E is moved to E'. It is important to note that this procedure does not find the smallest consistent subset from E. The CNN is sensitive to noise and noisy examples are likely to be misclassified as many of them will be added to the training set.

Tomek links [25] removes both noise and borderline examples. Tomek links are pairs of instances of di?erent classes that have each other as their nearest neighbors. Given two examples x and y belonging to different classes, and $d(x, y)$ the distance between x and y, a (x, y) pair is called a Tomek link if there is not an example z such that $d(x, z) < d(x, y)$ or $d(y, z) < d(x, y)$. If two examples form a Tomek link, then either one of these examples is noise or both examples are border-line. As an under-sampling method, only examples belonging to the majority class are eliminated. The major drawback of Tomek Link undersampling is that this method can discard potentially useful data that could be important for the induction process. This method has an higher order computational complexity and will run slower than other algorithms.

While the previous methods are intelligent down sampling techniques, SMOTE is an over-sampling method that produces new synthetic minority class examples. SMOTE [7] forms new minority class examples by interpolating between several minority class examples that lie together in "feature space" rather than "data space". For each minority class example, this algorithm introduces synthetic examples along the line segments joining any/all of the k minority class nearest neighbors (in this work k is equal to 3). Synthetic samples are produced taking the difference between the feature vector (sample) under consideration and its nearest neighbors. The difference is multiplied by a random number between 0 and 1 and added to the feature vector under consideration.

6 Evaluation Issues

One of the most used metric is the Error Rate, defined as 1.0- (True Positive+True Negative)/(True Positive-False Positive+False Negative+True Negative). However using this metric implies that the class distribution is known and fixed, an assumption that does not hold in real world applications as the one proposed here. Moreover, Error Rate is biased to favor the majority class, making it a bad choice when evaluating the effects of class distribution. Other aspect against the use of Error Rate is that it considers different classification errors as equally important, and in domains such medical diagnosis, the error of diagnosing a sick patience as healthy is a fatal error while the contrary is considered a much less serious error. This means that a metric such as Error Rate is sensitive to class imbalance.

It is possible to derive metrics that are not sensitive to the skew of the data. In particular, four metrics are proposed in [4]:

- False Negative rate: F N /(T P +F N) - the percentage of positive examples misclassified as belonging to the negative class
- False Positive rate: F P /(F P +T N) - the percentage of negative examples misclassified as belonging to the positive class
- True Negative rate: T N /(F P +T N) - the percentage of negative examples correctly classified as belonging to the negative class
- False Positive rate: T P /(T P +F N) - the percentage of positive examples correctly classified as belonging to the positive class

A good classifier should try to minimize FN and FP rates, and maximize TN and TP rates. Unfortunately, there is a tradeoff between these two metrics, and in order to analyze this relationship ROC graphs are used. ROC graphs are two-dimensional graphs where TP rate is plotted on the Y axis and FP rate is plotted on the X axis. ROC graphs are consistent for a given problem even if the distribution of positive and negative instances is highly skewed.

It is important to notice that the lower left point (0, 0) represents the strategy of never issuing a positive classification: such a classifier produces no false positive errors but also gains no true positives. The opposite strategy, of unconditionally issuing positive classifications, is represented by the upper right point (1, 1).

In order to compare classifiers, it is possible to reduce a ROC curve to a scalar value representing the performance of the classifier. Area Under the ROC (AUC) is a portion of the area of the unit square. Its value will always be between 0 and 1. However, because random guessing produces the diagonal line between(0,0) and(1,1), which has an area of 0.5, no realistic classier should have an AUC less than 0.5. The AUC is equivalent to the Wilcoxon test of ranks and it is also related to Gini coefficient (for an exhaustive description of ROC and AUC in assessing machine learning algorithms see [5]) . In this work, we will use the AUC measure in order to assess the performance of classifiers. Furthermore, for each classifier, we present also the F-measure in order to compare our results to previous works in this area. F-measure is a combination of Recall and Precision metrics:

$F - measure = \frac{2*Precision*Recall}{(Precision+Recall)}$

7 Results and Discussion

In this section, we show the results obtained with the different learning algorithms and with the different sampling techniques used for both corpora. We also present results obtained using the original data set, which is the data set with the original imbalance. This result represents our base line against which results obtained with sampled data sets are to be compared with. Values in bold represent the best score for each classifier.

Tables 1 and 2 display the performance of the two classifiers using k-NN algorithm. In particular Table1 reports on the results of the most basic implementation of k-NN, that is with k equal to 1 (1-NN). In this case a test example is simply assigned to the class of its nearest neighbor. Table 2 displays results obtained by

1-NN				
	P	T	D	U
Sampling	F-m	AUC	F-m	AUC
Original	0.19	0.56	0.06	0.55
Dowsampling	0.62	0.57	**0.57**	0.55
Oversampling	0.36	0.55	0.18	0.52
SMOTE	**0.63**	**0.66**	0.40	**0.70**
CNN	0.23	0.52	0.56	0.54
Tomek	0.57	0.59	0.35	0.56

Table 1: *Results using **k-NN** algorithm with **k=1***

3-NN				
	P	T	D	U
Sampling	F-m	AUC	F-m	AUC
Original	0.17	0.57	0.20	0.51
Dowsampling	0.62	0.59	**0.59**	0.61
Oversampling	0.51	0.58	0.33	0.56
SMOTE	**0.66**	**0.70**	0.42	**0.74**
CNN	0.65	0.61	0.57	0.55
Tomek	0.64	0.66	0.28	0.63

Table 2: *Results using **k-NN** algorithm with **k=3***

C4.5				
	P	T	D	U
Sampling	F-m	AUC	F-m	AUC
Original	0.17	0.65	0.09	0.49
Dowsampling	0.58	0.59	0.66	0.67
Oversampling	0.37	0.67	0.25	0.65
SMOTE	**0.77**	**0.87**	**0.81**	**0.91**
CNN	0.62	0.61	0.55	0.56
Tomek	0.63	0.60	0.58	0.63

Table 3: *Results using **C4.5** algorithm*

Random Forest				
	P	T	D	U
Sampling	F-m	AUC		
Original	0.13	0.65	0.02	0.56
Dowsampling	0.57	0.65	0.61	0.69
Oversampling	0.21	0.64	0.02	0.64
SMOTE	**0.75**	**0.94**	**0.77**	**0.96**
CNN	0.59	0.66	0.58	0.58
Tomek	0.65	0.59	0.61	0.73

Table 4: *Results using **Random Forest** algorithm*

a classifier using a k-NN algorithm with k equal to 3 (3-NN).

Regarding the results obtained with the algorithm 1-NN in Table 1, it is interesting to notice that, for the AUC metric, only the SMOTE sampling technique is able to significantly improve the base line for both corpora. For the Portuguese corpus there is an improvement of 10 points while for the Dutch corpus the improvement is even greater, reaching 15 points. The situation is slightly different for the F-measure. In this case, the best result is obtained by SMOTE for the Portuguese and by down sampling for Dutch. Results obtained with the 3-NN algorithm are very similar to those obtained with the 1-NN in terms of which sampling technique shows the greater improvements. It is worthwhile to notice that although the base lines for the above classifiers are very similar, they differ in the way they respond to the sampling techniques. In particular the 3-NN algorithm seems to take more advantage from the use of sampling, since it obtains better results in all the experiments and for both languages.

The results displayed in Table 3 refer to the best setting for the C4.5 classifier, where the tree was pruned using the C4.5s standard pruning procedure and no Laplace correction. Regarding Table 4, the classifier was built using 10 different trees. For both corpora SMOTE sampling method presents the best results in terms of AUC and F-measure, but in the case of Dutch the improvement regarding the base line was much greater in comparison with the improvement for Portuguese. Even if the base Iine for Dutch was worst at the end the it outperformed results obtained with the Portuguese corpus. The same observation holds for results present in Table 4.

Table 5 displays results obtained with a SVM classifier using a sigmoid kernel. The AUC base line for this classifier is very low, with a value below or equal to 0.5. With the use of sampling techniques the performance of this classifier is comparable to the 1-NN. Although SVM is a complex algorithm, it achieves a performance similar to the simplest algorithm used in this work, namely 1-NN. Furthermore it is the only classifier where the SMOTE does not show the best result, considering either AUC or F-measure.

The results in Table 6 refer to a Naïve Bayes classifier using normal distribution. As for the previous algorithm (except for SVM), the best results is obtained with the SMOTE technique, but there is a difference between the two corpora. For the Portuguese data set the base line is higher than for the other classifiers in terms of both metrics taken in consideration, but the improvements achieved with the use of sampling do not outperform the performance of other classifiers, namely C4.5 and Random Forest. On the other hand, for the Dutch data set the best results are obtained with Naïve Bayes even if the initial base line is similar to that obtained with 3-NNm atleast regarding F-measure.

In general for both the languages, the SMOTE sampling technique shows the best results in terms of AUC, followed by Tomek Link and Random oversampling. These results are comparable with those reported in the literature on imbalanced data sets in general. In a comprehensive study on the behavior of several methods for balancing training data, using 11

SVM				
	P	T	D	U
Sampling	F-m	AUC	F-m	AUC
Original	0.12	0.48	0.02	0.50
Dowsampling	**0.67**	**0.68**	**0.65**	**0.65**
Oversampling	0.61	0.59	0.60	0.64
SMOTE	0.60	0.60	0.32	0.59
CNN	0.59	0.57	0.61	0.59
Tomek	0.64	0.49	0.63	0.66

Table 5: *Results using **SVM** algorithm*

	Naïve	Bayes		
	P	T	D	U
Sampling	F-m	AUC		
Original	0.24	0.66	0.12	0.75
Dowsampling	0.62	0.62	0.70	0.72
Oversampling	0.67	0.68	0.68	0.75
SMOTE	**0.72**	**0.76**	**0.95**	**0.97**
CNN	0.64	0.63	0.66	0.69
Tomek	0.69	0.72	0.67	0.77

Table 6: *Results using **Naïve Bayes***

UCI data sets [2], Batista and colleagues [4] show that in most cases and with several data sets in different domains SMOTE and Random over-sampling are the most effective methods. In general, they lead to a rise in the AUC metric of few percentage points (1 to 4), when the base line was already high (more than 0.65), while where the base line was under this value the improvement was comparable to the one obtained in our work. In particular for the flag data set, they obtained an improvement of 34 percentage points.

Focusing on Natural Language applications [15] apply these methods to sentence boundary detection in speech, showing that SMOTE and down-sampling get the best results with an AUC of 0.89 (the base line being 0.80). However, they did not experiment intelligent down-sampling methods such as CNN or Tomek Link. Batista in [3] gets the best results in terms of AUC with an improvement of 4 percentage points on the original data set using a combination of SMOTE with Tomek link, followed by simple SMOTE, in a case study on automated annotation of keywords.

In our case the improvement regarding the original data set is between 10 and 29 percentage points, demonstrating how these methods can be effective in this application.

Regarding the comparison with other work in definition extraction, the improvement obtained on the F-measure, with the best result of 0.77 with C4.5 classifier, outperforms most of the systems using learning algorithms, confirming the importance of sampling techniques in supporting definition extraction tasks. [26], using the same corpus we used, reports on a F-measure of 0.73, obtained with a combination of syntactic rules and a Naïve Bayes classifiers for Dutch while [21], with a similar approach, but for the Polish language, obtain a F-measure of 0.35. Furthermore in all these works a combination of features are used in order to reach best results, while in this paper we only use bi-grams of POS as features. To conclude, our results are comparable with systems that represent the state of the art in the area, such as DEFINDER, which shows a F-mesure of 0.80.

8 Conclusions and Future Work

In this paper we have compared the performance of different learning algorithms and different sampling technique on a definition extraction task, using data sets in different language. Results presented show that this approach can be very effective in comparison to handcrafted rule to extract definitions, in terms of amount of time and performance. Furthermore techniques here presented are language and domain independent, making them a interesting resource in the field of Question Answering. Next steps in our researches will be integrate our classifier in a QA system in order to test this results in a much real world context.

[2] http://archive.ics.uci.edu/ml/

References

[1] M. Avelãs, A. Branco, R. D. Gaudio, and P. Martins. Supporting e-learning with language technology for portuguese. In *Proceedings of the International Conference on the Computational Processing of Portuguese (PROPOR2008)*. Springer, 2008.

[2] G. Barnbrook. *Defining Language: a local grammar of definition sentences*. John Benjamins Publishing Company, 2002.

[3] G. E. A. P. A. Batista, A. L. C. Bazzan, and M. C. Monard. Balancing training data for automated annotation of keywords: a case study, 2003.

[4] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6(1):20–29, 2004.

[5] A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.

[6] X. Chang and Q. Zheng. Offline definition extraction using machine learning for knowledge-oriented question answering. In D.-S. Huang, L. Heutte, and M. Loog, editors, *ICIC (3)*, volume 2 of *Communications in Computer and Information Science*, pages 1286–1294. Springer, 2007.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[8] I. Fahmi and G. Bouma. Learning to identify definitions using syntactic feature. In R. Basili and A. Moschitti, editors, *Proceedings of the EACL workshop on Learning Structured Information in Natural Language Applications*, Trento, Italy, 2006.

[9] H. Gernot. Defining patterns in translation studies: Revisiting two classics of german translation. *Translationswissenschaft in Target*, 19(2):197–215, 2007.

[10] P. E. Hart. The condensed nearest neighbor rule (corresp.). *Information Theory, IEEE Transactions on*, 14(3):515–516, May 1968.

[11] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

[12] H. Joho and M. Sanderson. Retrieving descriptive phrases from large amounts of free text. In *Proceeding of hte 9th international conference on Information and knowledge management*, pages 180–186, 2000.

[13] J. Klavans and S. Muresan. Evaluation of the DEFINDER system for fully automatic glossary construction. In *Proceedings of the American Medical Informatics Association Symposium (AMIA 2001)*, 2001.

[14] L. Kobylinski and A. Przepiorkowski. Definition extraction with balanced random forests. In A. Ranta, editor, *GoTAL 2008*, pages 237–247, Gothenburg, 2008. Springer-Verlag Berlin Heidelberg.

[15] Y. Liu, N. V. Chawla, M. P. Harper, E. Shriberg, and A. Stolcke. A study in machine learning from imbalanced data for sentence boundary detection in speech. *Computer Speech & Language*, 20(4):468–494, 2006.

[16] V. Malais, P. Zweigenbaum, and B. Bachimont. Detecting semantic relations between terms in definitions. In *the 3rd edition of CompuTerm Workshop (CompuTerm 2004) at Coling 2004*, pages 55–62, 2004.

[17] S. Miliaraki and I. Androutsopoulos. Learning to identify single-snippet answer to definition questions. In *Proceeding of the 20th International Conference on Computational Linguistic (COLING 2004)*, pages 1360–1366, Geneva, Switzerland, 2004.

[18] I. N. and S. K. Xml, corpus encoding standard, document xces 0.2. Technical report, Department of Computer Science, Vassar College and Equipe Langue ed Dialogue, New York, USA and LORIA/CNRS, Vandouvre-les-Nancy, France, 2002.

[19] J. Person. The expression of definitions in specialised text: a corpus-based analysis. In M. Gellerstam, J. Jaborg, S. G. Malgren, K. Noren, L. Rogstrom, and C. Papmehl, editors, *7th Internation Congress on Lexicography (EURALEX 96)*, pages 817–824, Goteborg, Sweden, 1996.

[20] A. S. Pinto and D. Oliveira. Extraçcão de definições no Corpógrafo. Technical report, Faculdade de Letras da Universidade do Porto, 2004.

[21] A. Przepiorkowski, M. Marcinczuk, and L. Degorski. Noisy and imbalanced data: Machine learning or manual grammars? In *Text, Speech and Dialogue: 9th International Conference, TSD 2008*, Brno, Czech Republic, September 2008. Lecture Notes in Artificial Intelligence, Berlin, Springer-Verlag.

[22] H. Saggion. Identifying definitions in text collections for question answering. In *LREC 2004*, 2004.

[23] J. R. Silva. Shallow processing of Portuguese: From sentence chunking to nominal lemmatization. Master's thesis, Universidade de Lisboa, Faculdade de Ciências, 2007.

[24] E. Tjong, K. Sang, G. Bouma, and M. de Rijke. Developing offline strategies for answering medical questions. In *Proceedings of the AAAI-05 workshop on Question Answering in restricted domains*, pages 41–45, 2005.

[25] I. Tomek. Two modifications of cnn. *Systems, Man and Cybernetics, IEEE Transactions on*, 6(11):769–772, November 1976.

[26] E. Westerhout and P. Monachesi. Extraction of Dutch definitory contexts for elearning purposes. In *CLIN proceedings 2007*, 2007.

[27] I. H. Witten and E. Frank. *Data Mining: Pratical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann, 2005.

Extracting Sense-Disambiguated Example Sentences From Parallel Corpora

Gerard de Melo
Max Planck Institute for Informatics
Saarbrücken, Germany
demelo@mpi-inf.mpg.de

Gerhard Weikum
Max Planck Institute for Informatics
Saarbrücken, Germany
weikum@mpi-inf.mpg.de

Abstract

Example sentences provide an intuitive means of grasping the meaning of a word, and are frequently used to complement conventional word definitions. When a word has multiple meanings, it is useful to have example sentences for specific senses (and hence definitions) of that word rather than indiscriminately lumping all of them together. In this paper, we investigate to what extent such sense-specific example sentences can be extracted from parallel corpora using lexical knowledge bases for multiple languages as a sense index. We use word sense disambiguation heuristics and a cross-lingual measure of semantic similarity to link example sentences to specific word senses. From the sentences found for a given sense, an algorithm then selects a smaller subset that can be presented to end users, taking into account both representativeness and diversity. Preliminary results show that a precision of around 80% can be obtained for a reasonable number of word senses, and that the subset selection yields convincing results.

Keywords

Example Sentence Extraction, Parallel Corpora, Disambiguation, Lexical Databases

1 Introduction

Many dictionaries provide not only definitions but also short sentences that demonstrate how a given word is used in context. Linguists and average dictionary users alike appreciate genuine examples of a word being employed in a sentence.

Goal An example sentence for a word sense is any genuine sentence that contains that word being used in the respective sense. A set of example sentences may (1) allow the user to grasp a word's meaning, and (2) see in what circumstances a word would typically be used in practice.

The first aspect is relevant because traditional intensional word definitions may be too abstract or even confusing to users of a dictionary. Often, the meaning of a word can be determined from its context. In conjunction with conventional definitions, example sentences may allow users to verify whether they have correctly interpreted a definition.

The second aspect is relevant since example sentences may reveal possible contexts a word can be used in. For instance, synonymous words such as '*child*' and '*youngster*' can have the same meaning, yet be used in somewhat different contexts. Examples provide evidence of typical collocations and expressions, e.g. the word '*birth*' often occurs as in '*to give birth*' or '*birth rate*' (but not $*$'*to give nascence*' or $*$'*nascence rate*').

For this reason, dictionaries typically include not only conventional definitions, but also example sentences that convey additional information about the meaning of a word. These are often short, limited in number, and in some dictionaries elicited rather than genuine. Hence, retrieving further example sentences can be helpful for lexicographical purposes, or to make the meanings and use more clear to language learners and other laypeople. In modern digital dictionaries, the tight space constraints of print media no longer apply, and thus a larger number of example sentences can be presented to the user on demand.

Our aim is to automatically obtain a set of sense-disambiguated example sentences that are known to mention a specific sense of a word. For instance, for a polysemous word such as '*bat*', we would like to obtain a set of example sentences that refer to the animal sense (e.g. '*There were many bats flying out of the cave.*'), and, separately, a list of example sentences that mention the word in its sports sense (e.g. '*In professional baseball, only wooden bats are permitted.*').

When a user browses a digital dictionary or lexical database, the example sentences could then be provided together with the relevant definitions of the word. Even in digital media, however, more examples may be available than can initially be displayed. For this reason, a means of choosing a restricted set of particularly representative example sentences is an additional requirement.

Contribution Our approach consists of two major building blocks that address the two issues just described. The first step (Section 2) involves extracting the sense-disambiguated example sentences from a parallel corpus by harnessing cross-lingual information to aid in assigning sentences to word senses. The second step (Section 3) selects a limited set of particularly representative example sentences for each word sense, using an algorithm that assesses the contributions made by individual sentences. We provide preliminary experimental results in Section 5.

Workshop On Definition Extraction 2009 - Borovets, Bulgaria, pages 40–46,

2 Example Extraction

In the example extraction step, we connect sentences from a corpus to word senses in a given sense inventory whenever we are sufficiently confident that the sentence is an example of the corresponding word being used in the respective sense.

Conventional word sense disambiguation heuristics could be used to determine word senses for a monolingual text, and then the sentences in that text could be linked to the respective senses. Unfortunately, even the most sophisticated all-words disambiguation techniques are currently not reliable enough when a fine-grained sense inventory is used [14].

The intuition behind our method is that, given a parallel text that has been word aligned, we can jointly look at both versions of the text and determine the most likely senses of certain words with significantly greater accuracy than for any single version of the text. After word alignment, we independently apply word sense disambiguation heuristics for each of the languages to obtain ranked lists of senses for each word. One then analyses to what degree the ranked lists for aligned words overlap. In many cases, this makes it possible to infer the sense of a word much more reliably than with conventional disambiguation heuristics. In such a case, we can use the respective sentence in which it occurs as an example sentence for that sense.

Lexical Alignment In the past, parallel corpora had been rather difficult to obtain. This has changed with the increasing multilinguality of the Web as well as the greater demand for such resources resulting from the rise of statistical machine translation. Resnik and Smith [15] showed that the Web can be mined to obtain parallel corpora, while Tiedemann [21] built such corpora from sources such as movie subtitles and manuals of open source software.

To compare the senses of words in both versions of a text, such parallel corpora first need to be word-aligned. This means that occurrences of terms (individual words or possibly lexicalized multi-word expressions) in one language need to be connected to the corresponding occurrences of semantically equivalent terms in the document for the other language.

This is usually accomplished by first aligning sentences, and then using global cooccurrence-based statistics to connect words of two corresponding sentences. Superficial similarities between words and part-of-speech information provide additional clues. We rely on pre-existing tools to perform this alignment, as will be explained in Section 5.

Disambiguation An important prerequisite for our approach is the existence of a word sense database. This resource must provide a fairly complete listing of word senses for a given word in any of the languages involved. We use the WordNet lexical database for the English language and the Spanish WordNet for the Spanish language (see Section 5).

Our system iterates over the sentences in the parallel corpus, simultaneously looking at two different languages a, b. Whenever an occurrence of a word t_a in a is aligned with a word t_b in b, and t_a is believed to be linked to a word sense s_a with a sufficiently high confidence score, we make the sentence where t_a was found an example sentence of s_a.

The confidence score is assigned as follows:

$$\mathrm{score}(t_\mathrm{a}, s_\mathrm{a}) = \mathrm{wsd}(t_\mathrm{a}, s_\mathrm{a}) \frac{\sigma(t_\mathrm{a}, s_\mathrm{a}) \mathrm{csim}(t_\mathrm{b}, s_\mathrm{a})}{\sum_{s' \in \sigma(t_\mathrm{a})} \sigma(t_\mathrm{a}, s') \mathrm{csim}(t_\mathrm{b}, s')}$$

The auxiliary function $\sigma(t)$ yields the set of all senses associated with t in the sense inventory, and $\sigma(t, s)$ is the corresponding indicator function ($\sigma(t, s) = 1$ if $s \in \sigma(t)$ and 0 otherwise).

In practice, looking up the possible senses of a word requires a morphological analysis to obtain lemmatized forms of words and determine their part-of-speech. We also rely on a look-ahead window to detect multi-word expressions occurring in the text that have their own sense identifier in the sense knowledge base.

The function $\mathrm{csim}(t_\mathrm{b}, s_\mathrm{a})$ measures the cross-lingual similarity between the likely senses of a term t_b in language b and a specific sense s_a for the word from language a:

$$\mathrm{csim}(t_\mathrm{b}, s_\mathrm{a}) = \sum_{s_\mathrm{b} \in \sigma(t_\mathrm{b})} \mathrm{sim}(s_\mathrm{a}, s_\mathrm{b}) \, \mathrm{wsd}(t_\mathrm{b}, s_\mathrm{b}) \qquad (1)$$

These functions build on a monolingual word sense disambiguation function $\mathrm{wsd}(t, s)$ and a sense similarity measure $\mathrm{sim}(s_1, s_2)$.

Monolingual Word Sense Disambiguation The $\mathrm{wsd}(t, s)$ function provides an initial monolingual disambiguation by measuring the similarity between the context of t in the corpus and a similar contextual string created for the sense s. For the former we use the current sentence being disambiguated (which contains t). The latter is created by concatenating glosses and terms associated with the sense s itself or with senses s' directly related via hyponymy, holonymy, derivation, or instance relations, or via up to 2 levels of hypernymy. These context strings are stemmed using the Porter algorithm [13], and feature vectors $\mathbf{v}(t)$, $\mathbf{v}(s)$ with term frequency values are created based on the bag-of-words vector space model. The result is then computed as

$$\mathrm{wsd}(t, s) = \sigma(t, s) \left(\alpha + \frac{\mathbf{v}(s)^T \mathbf{v}(t)}{||\mathbf{v}(s)|| \; ||\mathbf{v}(t)||} \right) \qquad (2)$$

Unlike standard word sense disambiguation setups, we prefer obtaining a weighted set of multiple possibly relevant senses rather than just the sense with the highest confidence score. We use α as a smoothing parameter: For higher values of α, the function tends towards a uniform distribution of scores among the relevant senses, i.e. among those with $\sigma(t, s) = 1$.

Semantic Similarity For the semantic similarity measure, we do not rely on generic measures of semantic relatedness often described in the literature [1]. The purpose of this measure here is to identify only word senses that are identical or nearly identical (e.g.

the senses for '*house*' and '*home*') rather than arbitrary forms of association (e.g. between '*house*' and '*door*').

We use the following relatedness measure:

$$\text{sim}(s_1, s_2) = \begin{cases} 1 & s_1 = s_2 \\ 1 & s_1, s_2 \text{ in near-synonymy relationship} \\ 1 & s_1, s_2 \text{ in hypernymy relationship} \\ 1 & s_1, s_2 \text{ in hyponymy relationship} \\ 0 & \text{otherwise} \end{cases}$$

The relational information between senses used here is provided by WordNet.

3 Example Selection

For computational applications, obtaining a repository of perhaps several hundred or even thousand examples for a single word sense can be useful. When displaying examples to human users, it is often better to provide a limited selection at first. The challenge then is deciding which sentences to choose.

We assume there is a space constraint in form of a limit k on the number of sentences that shall be presented to the user. Given a possibly large number of example sentences for a specific word sense, we must choose up to k example sentences that showcase typical contextual collocations and thereby aid the user in discerning the meaning and use of a term.

Assets Each example sentence can be thought of as having certain assets in this respect. For example, for the financial sense of the word '*account*', the fact that an example sentence contains the bigram '*bank account*' could be considered an asset. Another sentence may contain the commonly used expression '*open an account*'.

Our approach looks at 7 different sets of assets (in our case, neighbourhood n-grams) for each example sentence x associated with a word sense.

- $A^1_m(x)$: the original unigram word occurrences for which the example is provided, e.g. '*account*' or '*accounts*' (note that there might be different word forms, and additionally, in WordNet, multiple synonymous words can in fact be associated with a single word sense identifier)
- $A^3_m(x)$: word 3-grams incorporating a preceding and a following word, e.g. '*bank account number*'
- $A^2_p(x)$: word 2-grams incorporating previous words, e.g. '*bank account*'
- $A^3_p(x)$: word 3-grams incorporating previous words, e.g. '*open an account*'
- $A^2_f(x)$: word 2-grams incorporating following words, e.g. '*account manager*'
- $A^3_f(x)$: word 3-grams incorporating following words, e.g. '*account number is*'
- $A^*_m(x)$: the entire sentence

For each of these n-gram sets A^1_m, A^3_m, A^2_p, etc., we also consider the corresponding counter function a^1_m, a^3_m, a^2_p, etc., that counts how often the n-gram occurs in the example sentence in the respective relative position. Usually, this will either be 0 or 1, though an example sentence may also contain multiple occurrences of the word being described, so higher values do occur. Note that in the above use of the words *unigram* and *n-gram*, if the original word being described is a multi-word-expression, it is only counted as one word, e.g. when considering examples for the multi-word expression '*bank account*' instead of just '*account*', the sequence '*opening a bank account*' would be considered a 3-gram.

Our aim will be to choose example sentences that provide representative examples of each of these n-gram sets, so each asset will be given a weight. $A^*_m(x)$, which contains the entire sentence, is a special case where we define $w(a)$ for $a \in A^*_m(x)$ to be the cosine similarity with the gloss context string, as for the word sense disambiguation in Section 2. These weights bias our selection towards example sentences that more clearly reflect the meaning of the word. Apart from this, each n-gram is given a weight based on its relative frequency within the set. For instance, with respect to A^3_p, a frequent expressions like '*open an account*' should receive a much higher weight than '*Peter's chequing account*'. For an n-gram a in the set $A^1_m(a)$, we assign a weight $w(a) = \frac{a^1_m(x,a)}{\sum_{i=1}^{n} a(x_i,a)}$, and equivalently for the other n-gram asset sets $A^3_m(x)$, $A^2_p(x)$, etc.

Objective Of course, at this point one could simply select the top k sentences with respect to the total weight of the n-grams they have as assets. Such an approach however is likely to lead to a very homogenous result set: n-grams with a high weight occur in many sentences, and hence could easily dominate the ranking.

Instead, we define the goal as follows: Given a set of assets A (in our case, n-grams), a set of items $X = \{x_1, \ldots, x_n\}$ (in our case, example sentences), each associated with specific assets $A(x_i) \subseteq A$ (in our case, the union of n-grams returned by A^1_m, A^3_m, A^2_p, etc.), and a limit k, the goal is to choose a set C of items with cardinality $|C| < k$ such that the total weight of the assets

$$\sum_{a \in \bigcup_{x \in C} A(x)} w(a) \tag{3}$$

is maximized.

While this formalization aims at ensuring that items with highly weighted assets occur in the example set, e.g. a sentence containing '*open an account*', it also enforces a certain level of diversity. The latter is achieved by counting the weight of each asset only once, thus if one sentence includes '*open an account*', then there is no direct benefit for including a second sentence with that same n-gram.

The goal can equivalently be expressed in an integer linear program formalization as follows. Define

$$a'(x_i, a) = \begin{cases} 1 & a \in A(x_i) \\ 0 & \text{otherwise.} \end{cases}$$

Our objective is then:

$$\begin{aligned} \text{maximize} \quad & \sum_a c_a w(a) \\ \text{s.t.} \quad & c_a \leq c_{x_1} a'(x_1, a) + \cdots + c_{x_n} a'(x_n, a) \\ & c_{x_1} + \cdots + c_{x_n} \leq k \\ & c_a, c_{x_i} \in \{0, 1\} \end{aligned}$$

This means that we wish to maximize the weight of the assets (n-grams) with $c_a = 1$, where c_a can only be 1 if an appropriate $c_{x_i} = 1$, i.e. an appropriate item (example sentence) x_i is chosen for the result set.

We use a greedy heuristic to find solutions, since the problem is NP-hard.

Proof. We prove the NP-hardness by reducing the NP-hard vertex cover problem to our setting. Given a graph $G = (V, E)$ and a positive integer k, the vertex cover problem consists in determining whether a set of vertices C of size at most k exists, such that each $e \in E$ is incident to at least one $v \in C$. Now set $n = |V|$ and define the items $x_0, \ldots, x_n$ to be the vertices $v \in V$. Further, define $A = E$ as the set of assets and $A(x_i)$ as the set of edges incident to x_i. Give these edges uniform weights $w(e) = 1$. Having determined k items that maximize Equation 3, we can then simply test whether the score is equal to $|E|$. If it is, then obviously there exists a set of at most k vertices such that every edge $e \in E$ is covered. If not, then no vertex cover with at most k vertices can exist, because otherwise we could choose that vertex cover as the set of items and obtain a higher objective score (since more edges would be covered). Hence, any vertex cover problem could be answered using an exact algorithm for our problem setting. □

Approach The algorithm we use (Algorithm 3.1) relies on a simple greedy heuristic. It repeatedly chooses the highest-ranked sentence $x \in X$ given the current asset weights w, then resets the weights $w(a)$ of all assets $a \in A(x)$ to zero to ensure that they are no longer considered when choosing further sentences. Ties can be broken arbitrarily (in practice, we first compare the disambiguation scores from Section 2 and choose the highest one).

Algorithm 3.1 Sentence Selection algorithm

```
1: procedure SELECT(X, k, w)
2:     C ← ∅
3:     while |C| < k ∧ |X| > 0 do
4:         x ← argmax_{x∈X\C} Σ_{a∈A(x)} w(a)
5:         C ← C ∪ {x}
6:         for all a ∈ A(x) do
7:             w(a) ← 0
8:     return C
```

Prior to running the algorithm, an additional filtering may be used. For instance, one may filter out examples that are too long or too short (e.g. incomplete phrases or headlines and titles). One could also allow hiding sentences with possibly offensive or vulgar language.

If the number of example sentences is too large to do a linear scan of all sentences (e.g. in the case of highly frequent words such as conjunctions), we may also choose to let the algorithm run on a smaller random sample $X' \subset X$ of sentences as input.

A useful feature of this greedy algorithm is that it allows emitting a ranked list of entities. Having run the algorithm for a large k, perhaps even $k = \infty$, we can easily obtain the respective output for any $k' < k$ simply by pruning the ranked list generated for k. This can be very useful for interactive user interfaces.

4 Related Work

Several means of generating example sentences for word senses have been proposed. Shinnou et al. [19] extract example sentences for a word from a corpus and attempt to distinguish senses by passing human-labelled sentences as input to a clustering algorithm. This method requires significant human involvement and unlike our approach does not disambiguate senses with respect to a specific sense inventory.

Chklovski and Mihalcea [2] presented a Web interface that asks Web users to tag sentences with the correct word sense and relies on active learning methods to select sentences that are hard to tag automatically.

A different approach suggested by Mihalcea [10] finds example sentences by using a set of seed expressions to create appropriate queries to Web search engines. For example, for the fibre optic channel sense of word '*channel*', appropriate queries would be '*optical fiber channel*', '*channel telephone*', '*transmission channel*'. This method works well when such multi-word constructions can be constructed and could be used to complement our approach.

Another more recent approach [11] clusters words based on a dependency parse of a monolingual corpus. This means that for each word a set of similar words is available. One then tries to match example sentences from the corpus with example sentences already given in WordNet, taking into account the word similarities.

Our approach uses a different strategy by relying on parallel corpora. The intuition that lexical ambiguities in parallel corpora can be resolved more easily has been used by a number of works on word sense disambiguation. Dagan et al. [3] provided an initial linguistic analysis of this hypothesis. Several studies [9, 5, etc.] then implemented this idea in word sense disambiguation algorithms. These approaches are similar to our work. They use simple heuristics on parallel corpora to arrive at sense-labelled data that can then be used *for* word sense disambiguation, while our approach relies on a word sense heuristic to create example sentences from a parallel corpus.

With regards to the challenge of selecting the most valuable examples, Fujii et al. [8] proposed a method for choosing example sentences for word sense disambiguation systems. Unlike our approach, which aims at representative examples for end users, their approach aims at examples likely to be useful for training a disambiguation system. Their proposal selects example sentences that are hard to classify automatically due to the associated uncertainty, so particularly clear examples of a word's use are in fact less likely to get elected. Rychly et al. [17] presented a semi-supervised selection system that learns scores based on combinations of weak classifiers. These classifiers rely on features

Corpus	Covered Senses	Example Sentences	Accuracy (Wilson interval)
OpenSubtitles English-Spanish	13,559	117,078	0.815 ± 0.081
OpenSubtitles Spanish-English	8,833	113,018	0.798 ± 0.090
OpenOffice.org English-Spanish	1,341	13,295	0.803 ± 0.081
OpenOffice.org Spanish-English	932	11,181	0.793 ± 0.087

Table 1: *Number and Accuracy of sense-disambiguated example sentences*

such as word lists, sentence/word length, keyword position, etc. Since the system does not take into account diversity when generating a selection, it would be interesting to combine our algorithm with the scores from their classifiers as additional assets.

5 Results

We conducted preliminary experiments on multiple corpora to evaluate the usefulness of our approach.

5.1 Resources

In terms of parallel corpora, we relied on parts of the OPUS collection [21], in particular the OpenSubtitles [22] and the OpenOffice.org corpora. We made use of GIZA++ [12] and Uplug [20] to produce the word alignments for these corpora. Additionally, we evaluated example sentence selection for undisambiguated sentences using a subset of the Reuters RCV1 corpus [16], consisting of 39,351 documents.

The following lexical knowledge bases were used to build up the sense inventory:

- The original Princeton WordNet 3.0 [7] for the English language.
- The Spanish WordNet jointly developed by three research groups in Spain [6]. Since it was created in alignment with WordNet 1.6, we applied sense mappings [4] to obtain sense identifiers aligned with the version 3.0 of WordNet.

When linking words in the corpus to this inventory, the TreeTagger [18] was used for morphological analysis.

5.2 Experiments

We generated sense-disambiguated example sentences for several setups, and evaluated random samples by assessing whether or not the word was indeed used in the sense determined by our method. The results were generalized using Wilson score intervals, and are presented in Table 1. The smoothing parameter α from Section 2 was set to 0.3. In Table 2, we provide a few anecdotic examples of the output.

In general, this approach yields high-quality example sentences compared to current systems for monolingual text [14]. Automatic word alignment is known to be error-prone, and many heuristics have been proposed to mitigate the effects of this, e.g. aligning in both directions and then intersecting the alignment. In our setting, incorrect alignments are unlikely to lead to incorrect example sentences. This is because two erroneously aligned words in most cases have very different meanings and hence are unlikely to share a semantically similar word sense.

The main cause of the inaccuracies we encountered instead turned out to be the sense inventory's incompleteness. For instance, when an English word has multiple senses shared by the aligned Spanish word, but the sense inventory only lists one of those senses for the Spanish word, our method would lead us to believe that that sense is the right one with high certainty. On a few occasions, incorrect output by the morphological analyser induced errors. For example, when the word '*shed*' was labelled a verb although it was used as a noun, the wrong sense was selected.

A drawback of our approach is that the number of word senses covered is limited. To some degree, this can be addressed by using larger corpora and more language combinations. A reasonably full level of coverage of the senses listed in WordNet would however likely also require relaxing the scoring functions to take into account also less obvious (and hence less reliable) input sentences.

We also applied the sentence selection approach described in Section 3. Table 3 provides ranked lists of example sentences created using Algorithm 3.1. It is clear that frequent collocations such as '*right side*', '*electrical current*', and '*when nightfall comes*' are given a high weight. We also see at least one example sentence wrongly associated with a sense ('*convey*'). Since the algorithm does not depend on sense-disambiguated example sentences, we additionally show sentences from the monolingual RCV1 corpus in Table 4. A larger number of example sentences is typically available here, so the algorithm suceeds even better at choosing sentences that highlight typical collocations, e.g. '*long term*', '*a long time*' for the word '*long*', or '*colonial rule*' and '*colonial power*' for '*colonial*'. The RCV1 corpus is strongly biased towards the financial domain, which is reflected in the example sentences chosen by the algorithm.

6 Conclusions and Future Work

We have presented a framework for extracting sense-disambiguated example sentences from parallel corpora and selecting limited numbers of sentences given space constraints.

In the future, we plan on exploiting alignments with additional languages by using additional versions of WordNet. This would be particularly useful for pairs of languages that are phylogenetically unrelated, as these are more likely to have different patterns of homonymy, and hence a word in one language is less likely to share more than one meaning with a word in the other language.

line (something, as a cord or rope, that is long and thin and flexible)	I got some fishing **line** if you want me to stitch that. Von Sefelt, get the stern **line**.
line (the descendants of one individual)	What **line** of kings do you descend from? My **line** has ended.
catch (catch up with and possibly overtake)	He's got 100 laps to **catch** Beau Brandenburg if he wants to become world champion. They won't **catch** up.
catch (grasp with the mind or develop an understanding of)	I didn't **catch** your name. Sorry, I didn't **catch** it.
talk (exchange thoughts, talk with)	Why don't we have a seat and **talk** it over. Okay I'll **talk** to you but one condition...
talk (use language)	But we'll be listening from the kitchen so **talk** loud. You spit when you **talk**.
opening (a ceremony accompanying the start of some enterprise)	We don't have much time until the **opening** day of Exhibition. What a disaster tomorrow is the **opening** ceremony!
opening (the first performance, as of a theatrical production)	It will be rehearsed in the morning ready for the **opening** tomorrow night. You ready for our big **opening** night?

Table 2: *Samples of Sense-Disambiguated Example Sentences from the OpenSubtitles Corpus (in some cases with multiple words for a single sense identifier)*

The approach could also be extended to *simultaneously* consider aligned sentences from more than two languages to harness example sentences when individual alignments of two languages do not provide enough information for a reliable disambiguation.

For sentence selection, one could consider investigating additional input information for the algorithm, e.g. sentence lengths.

References

[1] A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[2] T. Chklovski and R. Mihalcea. Building a sense tagged corpus with open mind word expert. In *Proc. ACL 2002 Workshop on Word Sense Disambiguation*, pages 116–122, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[3] I. Dagan and A. Itai. Two languages are more informative than one. In *Proc. ACL 1991*, pages 130–137, 1991.

[4] J. Daudé, L. Padro, and G. Rigau. Making wordnet mappings robust. In *Proc. 19th Congreso de la Sociedad Española para el Procesamiento del Lenguage Natural (SEPLN)*, Universidad de Alcalá de Henares. Madrid, Spain, 2003.

[5] M. Diab. An unsupervised method for multilingual word sense tagging using parallel corpora: a preliminary investigation. In *Proc. ACL 2000 Workshop on Word Senses and Multilinguality*, pages 1–9, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[6] J. Farreres, G. Rigau, and H. Rodríguez. Using wordnet for building wordnets. In *Proc. COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montral, Canada, 1998.

[7] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.

[8] A. Fujii, T. Tokunaga, K. Inui, and H. Tanaka. Selective sampling for example-based word sense disambiguation. *Computational Linguistics*, 24(4):573–597, 1998.

[9] W. A. Gale, K. W. Church, and D. Yarowsky. Using bilingual materials to develop word sense disambiguation methods. In *Proc. 4th International Conference on Theoretical and Methodological Issues in Machine Translation: Empiricist vs. Rationalist Methods in MT*, pages 101–112, Montreal, Canada, 1992.

[10] R. Mihalcea. Bootstrapping large sense tagged corpora. In *Proc. 3rd International Conference on Language Resources and Evaluation (LREC), Las Palmas*, 2002.

[11] B. Z. Nik Adilah Hanin and F. Fukumoto. Example-assignment to wordnet thesaurus based on clustering of similar words. *IPSJ SIG Notes*, 2008(46):59–64.

[12] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[13] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[14] S. Pradhan, E. Loper, D. Dligach, and M. Palmer. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proc. 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[15] P. Resnik and N. A. Smith. The web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380, 2003.

[16] Reuters. Reuters Corpus, vol. 1: English Language, 1996-08-20 to 1997-08-19, 2000.

[17] P. Rychly, M. Husak, A. Kilgarriff, M. Rundell, and K. McAdam. GDEX: automatically finding good dictionary examples in a corpus. In *Proc. XIII EURALEX International Congress*, Barcelona, Spain, 2008.

[18] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Intl. Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[19] H. Shinnou and M. Sasaki. Division of example sentences based on the meaning of a target word using semi-supervised clustering. In *Proc. 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, 2008.

[20] J. Tiedemann. Combining clues for word alignment. In *Proc. EACL 2003*, pages 339–346, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[21] J. Tiedemann. The OPUS corpus - parallel & free. In *Proc. 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 2004.

[22] J. Tiedemann. Improved sentence alignment for movie subtitles. In *Proc. RANLP 2007*, 2007.

being or located on or directed toward the side of the body to the east when facing north	1. In America we drive on the **right** side of the road. 2. I'll tie down your **right** arm so you can learn to throw a left. 3. If we wait from the **right** side, we have an advantage there.
put up with something or somebody unpleasant	1. You can't **stand** it can you? 2. You really think I can **tolerate** such an act? 3. No one can **stand** that harmonica all day long.
using or providing or producing or transmitting or operated by electricity	1. Not the **electric** chair. 2. Some **electrical** current circulating through my body. 3. Near as I can tell it's an **electrical** impulse.
take something or somebody with oneself somewhere	1. And they were kind enough to **take** me in here. 2. It **conveys** such a great feeling. 3. We interrupt this program to **bring** you a special news bulletin.
the time of day immediately following sunset	1. When **nightfall** comes go get dressed for the show. 2. You have until **dusk** to give yourselves up. 3. At **dusk** they return loaded with fish.

Table 3: *Example Sentence Rankings (OpenSubtitles Corpus)*

long	1. In the **long** term interest rate market, the yield of the key 182nd 10 year Japanese government bond (JGB) fell to 2.060 percent early on Tuesday, a record low for any benchmark 10-year JGB. 2. "The government and opposition have gambled away the last chance for a **long** time to prove they recognise the country's problems, and that they put the national good above their own power interests", news weekly Der Spiegel said. 3. As **long** as the index keeps hovering between 957 and 995, we will maintain our short term neutral recommendation.
colonial	1. Hong Kong came to the end of 156 years of British **colonial** rule on June 30 and is now an autonomous capitalist region of China, running all its own affairs except defence and diplomacy. 2. The letter was sent in error to the embassy of Portugal – the former **colonial** power in East Timor – and was neither returned nor forwarded to the Indonesian embassy. 3. Sino-British relations hit a snag when former Governor Chris Patten launched electoral reforms in the twilight years of **colonial** rule despite fierce opposition by Beijing.
purchase	1. Romania's State Ownership Fund (FPS), the country's main privatisation body, said on Wednesday it had accepted five bids for the **purchase** of a 50.98 percent stake in the largest local cement maker Romcim. 2. Grand Hotel Group said on Wednesday it has agreed to procure an option to **purchase** the remaining 50 percent of the Grand Hyatt complex in Melbourne from hotel developer and investor Lustig & Moar. 3. The **purchase** price for the business, which had 1996 calendar year sales of about $25 million, was not disclosed.
gold	1. Coach Ian Stacker said his team had hoped to meet the US in the **gold** medal play offs, but because of an early loss to Turkey the team did not get the draw they had counted on. 2. He said India's exports of **gold** and silver jewellery were worth $600 million annually against world trade of about $20 billion. 3. In the bullion market spot **gold** was quoted at $323.80/30 early compared to the London morning fix of $324.05 and the New York close Friday of $324.40/90.

Table 4: *Example Sentence Rankings (RCV1 Corpus)*

A Proposal for a Framework to Evaluate Feature Relevance for Terminographic Definitions

Selja Seppälä
University of Geneva, School of Translation and Interpretation (ETI)
Department of Multilingual Information Processing (TIM)

seppala2@etu.unige.ch

Abstract

In this paper, a terminological framework, both theoretical and methodological, backed by empirical data, is proposed in order to highlight the particular questions to which attention should be paid when conceiving an evaluation scheme for definition extraction (DE) in terminology. The premise is that not just any information is relevant to defining a given concept in a given expert domain. Therefore, evaluation guidelines applicable to DE should integrate some understanding of what is relevant for terminographic definitions and in which cases. This, in turn, requires some understanding of the mechanisms of feature selection. An explanatory hypothesis of feature relevance is then put forward and one of its aspects examined, to see to what extent the example considered may serve as a relevance referential. To conclude, a few methodological proposals for automating the application of relevance tests are discussed. The overall objective is to explore ways of empirically testing broader theoretical hypotheses and principles that should orient the conception of general guidelines to evaluate DE for terminographic purposes.

Keywords

Terminology, terminographic definitions, evaluation guidelines, terminological theory, terminographic methodology, concepts, feature relevance

1 Introduction

Definition extraction (DE) evaluation in terminology may be seen as a task aimed at enhancing precision and reducing the noise generated, for example, by limited extraction algorithms, i.e. as a task consisting in separating information on a concept from other information (for example, on another concept). Thus considered, the task of evaluation would consist in assessing whether all the information about a concept (i.e. all the conceptual contexts in which it occurs) has been retrieved, and whether no extraneous or spurious information has been retrieved. Assuming that this conceptual context retrieval issue is settled and that we already have all the textual contexts relating to a concept we want to define, there is still another aspect to be evaluated: is it the case that all the information extracted on a given concept in a given specialized corpus is relevant to the definition of that concept in that expert domain. One could argue that since the corpus from which the information is extracted is a specialized one, all the extracted information on a concept is at least potentially defining. However, as we shall see, this is not always the case. How may it be possible, then, to decide what is (or may be) relevant to the definition of a concept and what is not? What is addressed here is, therefore, a more fundamental kind of evaluation concerning the relevance of the extracted information for terminographic definition writing.

In that perspective, we shall first show that what is extracted is not necessarily a definition, basing our argument on terminological and terminographic frameworks as well as on an empirical study.

This background implies several questions which ought to be considered when designing an evaluation scheme applicable to extracted information and its use for terminographic definitions. Some hypotheses concerning the elements against which the extracted information may be evaluated are proposed and examined, as are methodological approaches to answering the questions thus raised, therefore providing empirical grounds for an evaluation. The main focus of this paper is therefore highlighting various factors that should be considered in evaluating the relevance of extracted information.

2 Background

2.1 Theoretical background

2.1.1 Relation between concepts and definitions

Facts and objects have innumerable properties, some of which are expressed in conceptual features, which are considered as more or less extended units of information. Not all of the features are of interest for experts when they form a concept encompassing a particular extension (facts or objects) in their expert domain; only *salient features* (F_S), as opposed to *latent features* (F_L), are. The latter are features associated to an extension but generally not expressed as such in human dictionaries. Latent features are often implied by and inherited through other features, such as the kind of entity and the high level properties associated with those entities. The latent features might never-

Workshop On Definition Extraction 2009 - Borovets, Bulgaria, pages 47–53,

theless be important in natural language processing lexicons or ontologies, for example for use in applications capable of drawing inferences. However, they are not expressed in terminological dictionaries, therefore they are not considered salient in that case. Latent features may also be features relating to the extension that are possessed by individuals as part of their background knowledge, but are not of interest to the domain under consideration. For instance, the fact that a *container* is *used to promote a brand* is something one may know about that object, but which is totally irrelevant in the domain of *waste management*, where what matters are the main functions of the object in that domain, such as *conditioning, transportation and storage of goods*, or the fact that *they are a large part of waste* and that *they have to be valorized by industrials themselves*. These latter features are thus considered salient.

Furthermore, not just any salient feature forming a specialized concept is of interest in defining that concept; only *relevant features* (F_R) are. In the previous example, only the main functions of a *container* and the information relative to its *valorization* are relevant in that particular domain. Thus, a definition is a set of relevant features which correspond to a subset of salient ones, or in fact often, as will be shown, to a set of potentially relevant ones (F_{PR}), i.e. a set of features that could each be perfectly relevant to a definition of the concept, but that are not necessarily selected to play a part in the definition.

2.1.2 Concepts and definitions in terminological dictionaries

This theoretical background should adequately account for the way in which conceptual information is conveyed in terminological dictionaries or databases, which gather specialized knowledge (*concepts*) by means of dictionary entries (*terminological records*). These are, indeed, composed of different fields corresponding to different kinds of information relating to the concept[1] —mainly term(s), definition, field code(s), encyclopedic note(s) and illustration(s). Each field expresses at least some salient feature(s) of the concept through linguistic or other means (symbols, schemas, illustrations, films, etc.). The definition expresses the relevant features of the concept.

The logical conclusion of the theoretical framework is that not just any feature (piece of information) is relevant to define a given concept in a given expert domain.

2.2 Methodological background

This conclusion has an impact on terminological methodology which shows in the terminographic practice of definition writing: to write a definition, one extracts from a specialized corpus all the salient information on the concept to be defined. After identification of the potentially relevant features among the extracted data, a further selection may be done. The resulting relevant features are then compiled in a single definition so as to express them in a single informative sentence.

2.3 Implications of the background

The present theoretical and methodological framework implies that definitions express only features of a concept that are relevant in a given context[2]. This, in turn, implies that a distinction must be made between theories of concepts and theories of definitions.

A further implication of this background is that the information extracted from corpora is not necessarily defining, let alone making up a full definition, although it may sometimes be the case (only 11 cases out of 56 analyzed concepts); the information extracted mostly corresponds to some feature of the concept (wether potentially relevant, salient or even latent) —out of 380 identified non redundant features, 242 were F_{PR}, 125 F_S and 13 F_L—, which in some rare cases correspond to elements of its extension (13/380 features).

3 Questions to be considered for DE evaluation

The last paragraph raises several questions pertinent to the design of schemes for the evaluation of information extraction for definitions in a terminological context.

3.1 First question: What kind of relevant information?

What kind of information is relevant and relative to what? Is there a general (universal) relevance rule that would be applicable to all possible cases, whatever the concept or the domain? Terminological concepts are often considered to be functional concepts, therefore defined in terms of a specific function. However, empirical studies show that this is not always the case (see for example [15]). Given the results of empirical findings, it appears that there is no such general relevance rule and that, in order to be able to evaluate the relevance of the extracted conceptual information in terms of their conceptual content, one ought to have an idea of what is relevant in particular cases.

The hypothesis proposed to address this question is that feature relevance depends on:

- the conceptual category of the defined entity (`ABSTRACT`, `INANIMATE`, `ANIMATE`, `EVENT`, etc.) and
- the type of expert domain to which the concept belongs [2, 9, 10, 11, 12].

These hypotheses build on the findings of two domains: On the one hand, findings in lexical semantics show that different types of entities imply different types of argument structures; on the other hand, research

[1] Terminological dictionary entries also contain linguistic information, i.e. on linguistic properties and behaviour of terms (spelling variants, phraseology, etc.), but these are not of interest here.

[2] It also follows that the term "concept" in terminology does not refer exactly to the same "concept" as in other domains, where it often refers to a wider concept, for instance one which would encompass all the possible features associated with a given extension by a given individual. The terminological "concept" nevertheless stays in line with the latter.

in cognitive science has shown that feature salience depends on the kind of entity considered and/or the type of theory.

This first question may therefore be answered empirically —in a manner proposed by [15], following [3, 13, 14]— by studying the internal structure of definitions in terms of the conceptual relations (such as `FUNCTION`, `PART`, `CAUSE`, `CONSEQUENCE`, etc.) that are conveyed by the features expressed in definitions[3]. Thus, an evaluation scheme should use the observed results (annotated genus-specific or extension patterns) to assess the relevance of the extracted information relative to the kind of entity defined and the kind of domain type it belongs to.

However, answering this first question is not sufficient to obtain a complete evaluation scheme: the method only describes what is deemed relevant in the cases studied (even though the results are generalized through proper statistical methods); it does not provide any understanding of why those results are observed. The results do not say anything about which other salient or potentially relevant features were discarded as non relevant or, if deemed potentially relevant, why they were excluded. They do not tell anything about relevance conditions. To put this differently: the definitions studied could have contained a larger number of features deemed to be relevant. For one reason or another, some potentially relevant features were excluded. We need to understand the reasons for the exclusion in order to be able to understand feature selection in definitions. Only then can we hope to evaluate the relevance of the extracted conceptual information for definition writing in terminology.

3.2 Second question: What relevance conditions?

Obviously, this raises another question: what are the relevance conditions? This question, in fact, implies two interrelated subquestions: what are the principles guiding the selection of relevant features:

> (2a) How to decide if a feature may be relevant, i.e. among salient features, how to distinguish between a defining information element and a non defining one, and
>
> (2b) what are the principles guiding the selection of relevant features to be introduced in the definition from amongst a larger set of potentially relevant features?

These sub-questions are interrelated in the sense that possible answers to one could well apply to both, as will be shown later. It is important, however, to make a distinction between these two questions, and to try to answer them separately. As to how to proceed to resolve them within this framework, an explanatory proposal can be put forward (4.2), which should be tested to determine which of its particular hypotheses may prove useful in answering the questions and, thus, to serve as a basis for feature evaluation. That is what will be illustrated in the following sections, by focusing on possible ways to answer question 2b.

[3] This question won't be addressed here.

3.3 Summary of the issues

To conclude this preliminary part, a visual summary (Fig. 1) of the set of questions relating to feature selection for definitions and which ought to be considered in elaborating an evaluation scheme for DE in terminology is proposed.

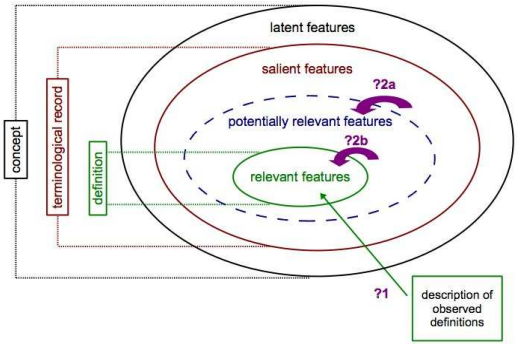


Fig. 1: *Questions related to feature selection*

This figure shows the nested sets of features, where the smallest set represents the relevant features actually expressed in a definition and which can be empirically studied in order to answer question 1. The arrow going from the set of salient features (expressed in a terminological record) to the possible set of potentially relevant features represents question 2a, and that going from the latter to the set of relevant features, question 2b. Again, 2a is a matter of distinguishing what is defining and what is not —for example, between a relevant feature and a so called encyclopedic information element—, and 2b a matter of feature selection amongst potentially defining features.

4 Proposals to answer question 2b

From here on, the focus will mainly be on some proposals concerning answers to question 2b, i.e. what are the relevance conditions or referentials, or what makes a feature relevant. We will first present a method that can be used to empirically select a feature as relevant (4.1). We will then focus on a more theoretical explanation (4.2).

4.1 Empirical method

On a purely empirical level, one may select relevant features (among a larger set of merely salient or potentially relevant features) by identifying repetitions in the retrieved information. Data analysis indeed shows that repetitions of a feature in the extracted contexts mostly correspond to relevant features (64 % of repeated features appeared in a definition). However, this method has some drawbacks: (1) It often requires that multiple contexts are retrieved for a given concept (73 % of the repeated features appeared in multiple contexts), and preferably from multiple sources. Yet, the analyzed data shows that multiple contexts were

found only for 50 % of the concepts. When relevant features are expressed within the same context, the repetition generally appears in the form of anaphoras or titles of paragraphs followed by the same information within the paragraph. (2) Automating the identification of redundant information may also prove difficult because of some divergences in the phrasing of the information. To be reliably used, this method should be further tested on large corpora. Another limitation of this method is that the solution proposed to the problem of deciding whether a feature is relevant is purely pragmatic (albeit probably very efficient); it lacks explanatory power. In the next section, therefore, we introduce some proposals which enjoy a more solid theoretical grounding.

4.2 An explanatory proposal

The following proposal to explain feature relevance combines well known factors that are generally thought to influence concept formation, and thereby feature selection: the kind of extension defined, the kind of theoretical context and individual background in which the concept is considered, and the communicative setting involved. At a more specific level, the proposal also takes into account feature characteristics, which are considered relative to the different dimensions, so as to individuate different hypotheses. Each hypothesis should be examined and tested in order to see to what extent the combination of a dimension with a given feature characteristic may serve as a particular relevance referential, for example with regard to automatic evaluation methods.

4.2.1 Constraining dimensions

One may consider (at least) three dimensions which should interact as relevance referentials in constraining feature selection for a definition: an *extensional dimension* where the objects of the extension are considered as such, independently of any context or domain, a *contextual dimension* encompassing *conceptual systems* and *individual backgrounds*, and a *communicative dimension*. These *dimensions* are taken to be higher order or more general referentials, and may be characterized by a set of attributes (such as the *type of object set* for the *extensional* dimension) having different values (for example, for *type of object set*, attributes distinguishing a *single object set* from a *multiple object set*). The attribute-value pairs may in turn be related to different types of feature characteristics. As presented in more detail below, the latter may also be described in terms of attributes (such as *feature coverage*) and values (like *universal feature* or *stereotypical feature* for *feature coverage*)[4].

4.2.2 Feature characteristics

Features may be characterized in several ways, also specified by means of attributes and values.

As far as content is concerned, a feature expresses some information, which may correspond to a *gradable* or an *non gradable property* (i.e. being more or less something) of the object(s) of the extension, and which may be described in terms of *conceptual category* (i.e. type of entity, for example `ABSTRACT`, `INANIMATE`, `ANIMATE` or `EVENT`) and *relation* (such as `FUNCTION`, `PART`, `CAUSE` or `CONSEQUENCE`). The expressed information may also contain part of the *extension of the concept*. Relative to the definition, it may correspond to the *genus* or to a *specific*, having either a *descriptive* or a *distinctive* function, and a *necessary* or *sufficient* status.

Relative to the object(s) of the extension, a feature may be characterized in terms of *feature coverage* as referring directly to a particular instance of the object(s) of the extension, in which case it may be called a *singular* or *individual feature*, or as a generalization covering either a certain percentage of the objects (*stereotypical feature*) or all of them (*universal feature*). Among the universal features, a further distinction may be drawn between those that belong to the extension alone (and are therefore *typical* or *distinctive*) and those that are shared with some other extension (and are therefore *non typical* or *non distinctive*[5]) [7]. The latter distinction (typical vs non typical) may also prove useful in determining wether a feature constitutes a distinguishing feature, setting a concept apart from other other concepts in the domain. A feature may also be described, again relative to the object(s) of the extension, as *intrinsic* or *extrinsic*, *essential* or *accidental*, and *necessary* or *non necessary*.

Finally, a feature may be described in terms of mental representations of three kinds: *theoretical* (T), *prototypical* (P) or *exemplar* (E) [4, 6], respectively a representation consisting in a causal or nomological understanding of a property (T), one associated with statistically typical features of a property (P), and one consisting in individual exemplars of a property as already encountered by a person (E).

5 Extensional constraints on feature selection

In the subsequent exploration of an answer to question 2b, concerning the selection of relevant features among potentially relevant ones, a method will be presented to illustrate how each of these feature characteristics may be examined and tested to see if it could be used as a relevance referential. In concrete terms, the method involves examining one single aspect or attribute of the extensional dimension, the type of object set, in relation to one single feature characteristic, feature coverage. i.e. *feature coverage*. It is an attempt to see (i) what this particular characteristic of features (4.2.2) can tell about feature selection relative to extensional constraints (5), if anything; and, (ii) what pragmatic factors (5.3) should be considered as further constraining feature selection, for example, how the other dimensions may enter into the selection process. It is not intended to give definitive answers; the focus is on exploring possible paths towards an answer given the proposed explanatory hypotheses, and on trying to identify possible methodologies for testing the adequacy of any proposed solution. This could

[4] From lack of space, this hypothesis will not be further elaborated on here.

[5] The term *typical* will be used to avoid confusion with a feature having a distinctive role within a conceptual system.

eventually lead to appropriate evaluation methods for extracted information. The overall objective is thus to have an understanding of where to search for solutions to the questions, which might help orienting evaluation methodologies (for example, automatic vs. human expert) for the feature selection aspect of definition extraction.

The particular focus on the coverage characteristic of features (*feature coverage*) and its relation to the type of extension, i.e. the *type of object set* defined (5.1), shows that their combination yields and licences different definition structures: *classical* or *by necessary and sufficient conditions* (where the sum of the features covers all the objects of the extension), *prototypical* (where the sum of the features does not cover all the objects of the extension) or *(semi-)encyclopedic* (where the sum of the features covers only the one object of the extension). The aim is to explore the relationship between this particular characteristic of features and definition structure.

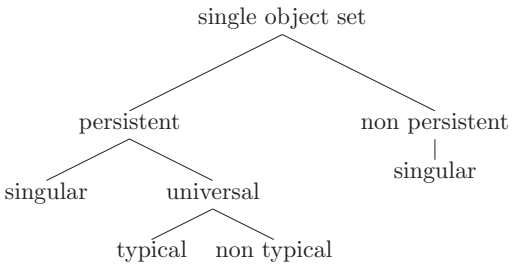
5.1 Types of object sets

An extension, considered as such, independently of any context or domain, can be described as corresponding to two different *types of object set* (one of the extensional dimension's attributes): a *single object set* and a *multiple object set*, which latter may contain a *homogenous* or a *heterogenous* set of objects. A *multiple object set* may also be described as a *closed set* in which the objects may be listed or an *open set* where it is not possible to list all the objects[6]. Single object extensions may consist in *persistent objects* (e.g. the sun, the earth) or in *non persistent objects* or *contingent objects* occurring at some particular location at a particular time (e.g. historical concepts, the swiss *Conseil fédéral*).

5.2 Correlations between feature coverage and definition structure

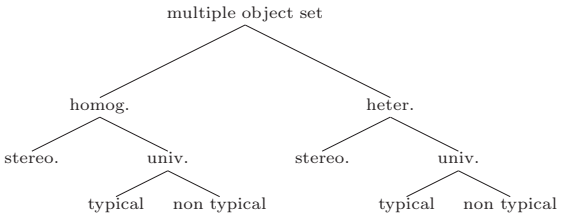
These different types of object sets or extensions may tend to correlate with different types of feature coverage in the following way.

For extensions composed of a single object:



[6] The distinction between open and closed set may prove useful as far as extensional definitions are concerned, but should not present any difference for the later argument here.

For extensions composed of multiple objects:



What definition structures are possible in what cases can now be deduced from the correlations between types of object sets and feature coverage, coupled with the knowledge that classical definitions may only use universal distinctive features whilst encyclopedic definitions may only use singular features. From these correlations, the constraint of the extensional dimension on feature coverage and the resulting definition structure may be expressed in the form of the following conditional rule:

```
<DEFINITION STRUCTURE>
licensed if
<TYPE OF EXTENSION and
FEATURE COVERAGE>
```

The extension and feature conditions for each definition structure may now be listed as follows:

```
CLASSICAL
if
<SINGLE PERSISTENT OBJECT and
UNIVERSAL TYPICAL FEATURE>
or
<HOMOGENOUS SET and
UNIVERSAL TYPICAL FEATURE>
or
<HETEROGENOUS SET and
UNIVERSAL TYPICAL FEATURE>

PROTOTYPICAL
if
<SINGLE PERSISTENT OBJECT and
UNIVERSAL NON TYPICAL FEATURE>
or
<HOMOGENOUS SET and
STEREOTYPICAL FEATURE>
or
<HOMOGENOUS SET and
UNIVERSAL NON TYPICAL FEATURE>
or
<HETEROGENOUS SET and
STEREOTYPICAL FEATURE>
or
<HETEROGENOUS SET and
UNIVERSAL NON TYPICAL FEATURE>

ENCYCLOPEDIC
if
<SINGLE PERSISTENT OBJECT and
SINGULAR FEATURE>
or
<SINGLE NON PERSISTENT OBJECT and
SINGULAR FEATURE>
```

According to this proposal —and provided it is tenable—, the nature of the extension should be considered as one of the relevance conditions or referentials for relevant feature selection. It shows, for example, that multiple object sets do not licence encyclopedic definitions, and also that they licence both prototypical and classical structures. It shows, furthermore, that *single features*, which are in principle banned from terminographic definitions, are in some

circumstances perfectly relevant to constructing a definition of a concept[7]. A fact that is not accounted for either by the classical theory of definitions or by the prototypical theory.

These conclusions entail that only in some cases does the feature-extension combination correspond to a single definition structure, and thus, that the corresponding potentially relevant features will (or should) be selected as actually relevant. In the other cases, further constraints are necessary to decide which definition structure applies. There are, for example, cases where heterogenous sets of objects considered in a given context or domain are defined by means of a classical definition, implying the use of some universal feature. That would be the case for a set of heterogenous objects which have the same function in a given domain: the `FUNCTION` feature applies to all of the objects of the extension and is thus a universal feature, licensing a classical definition [5, 190]. Therefore, the extension alone, considered independently of any context, is not a sufficient relevance referential. Further pragmatic constraints may enter in the relevant feature selection decision.

5.3 Further pragmatic constraints on relevant feature selection

In a second step, some pragmatic factors that may further explain relevant feature selection are put forward. These factors are partly related to the other dimensions postulated as broader feature relevance referentials, that is the *contextual* and *communicative dimensions*.

5.3.1 Contextual and communicative constraints

The nature of the vocabulary used to express the features in a terminological record (for instance, in the term(s) referring to the concept) is partly dependent on the target audience and its background knowledge. This is because definitions do not function in isolation; they are always considered in conjunction with the other information expressed in the terminological record. Therefore, if the other fields express an item of information in an explicit manner (for instance, if the terms are deemed "transparent"), then —according to the concision principle— the definition may be relieved of those features that are already expressed elsewhere.

5.3.2 Methodological constraints

The type of terminographic work carried out may also determine which features should be expressed in another field instead of in the definition. For instance, in systematic terminography[8], some features that are potentially relevant may be expressed in the field codes for the sake of concision and to ease indexing and concept retrieval. In those cases, the feature which is considered as defining, and which could therefore be relevant, may only appear in the field code and is thus only indirectly inherited.

Some potentially relevant features may also be included in the genus' comprehension and are thus implicitly inherited [16, 31]. Those potentially relevant features may therefore be omitted from the definition.

6 Methodological proposals for DE evaluation

Now that a tentative (and partial) solution to the problem of selecting relevant features amongst a set of potentially relevant features has been examined, some equally tentative methods for automating the evaluation of information extracted from a corpus for use in constructing terminological definitions may be considered. The methodologies proposed should be tested in order to see whether they can in fact be automated or if the evaluation process needs to be carried out by a human expert. In any case, it should be noted that this empirical endeavor is a task whose accomplishment relies on linguistic factors. Therefore, to be applicable, each method should be associated with a set of linguistic features specifically devised for each language.

Considering that the relevance referentials examined are the coverage of a feature and the nature of the extension's object set, two methodological questions should be addressed:

1. How to account for feature coverage automatically, i.e. the number of objects in the extension?
2. How to account for the extension's nature automatically, i.e. the homogeneity or heterogeneity of the set of objects?

As far as *feature coverage* is concerned, we suggest identifying and testing linguistic patterns that could be matched with the three types of feature coverage. Thus, *universal features* might be searched out by looking for nomological expressions like "all N[...]" or "always found in[...]"; *stereotypical features* might be identified by looking for generalizing expressions like "generally[...]" as suggested by Pearson [8, 142–143] (where more patterns are proposed that may serve to determine the one or the other type of feature coverage or another), or expressions like "measuring between [...] and [...]", which express features that allow for some variation in the properties of the extension's objects[9]; and, finally, *singular features* might be found by identifying referential expressions, such as proper names.

[7] In this case, one could say that this proposition also answers question 2a in that it specifies particular cases where encyclopedic information is actually defining. However, wether just any single feature is defining (as opposed to non defining) is a different question. As an example, consider the difference between two features specifying the swiss political entity *Conseil fédéral*: *is swiss* is (or, at least, may be considered as) defining, but *was founded in the year [...]* is not.

[8] That is, terminographic work carried out systematically on all the terms of a domain or a subdomain, as opposed to punctual terminography, where the work is done on a term by term basis.

[9] This kind of expression is also a sign of gradability, which may not be incompatible with universal features. It could therefore be ambiguous.

In order to identify *multiple object sets*, one could look, following Carlson [1], for expressions like "are widespread", which only apply to kind predicates, thus exclude single object extensions. With respect to identifying the nature of the *multiple object extension*, an expression identified as a genus by way of a linguistic marker, for example, and followed by a disjunction would be a sign of a *heterogenous set*. However, this kind of judgement may prove difficult to automate. Expressions like "for example" followed by an enumeration may also indicate the heterogeneity of the extension. Indeed, in some cases, it is difficult even for a human annotator to determine whether the set is homogenous or heterogenous.

These methodological questions should be further examined in order to determine to what extent this particular hypothesis may be used as a reliable feature relevance referential, and to see if its evaluation is easily automated or if evaluation needs to be performed by a human judge, who may apply a wider range of tests to asses the relevance of a feature —of extracted information— with respect to the extension, while considering the exemplified attributes and values. She might not only rely on linguistic markers, but also use linguistic tests that require making more complex inferences, not to mention make use of tests that are based on her understanding and interpretation of the information in the extracted text, or on her background knowledge of the world.

7 Conclusion

On the grounds that not just any extracted information element is relevant to the definition of a given concept in a given expert domain, it was claimed that evaluation guidelines applicable to DE should integrate some understanding of what is relevant for terminographic definitions and in which cases. This, in turn, requires some understanding of the mechanisms of feature selection. The main purpose of this paper was, therefore, to narrow down questions and possible answers in order to determine where adequate solutions to the problem of relevant feature selection can be sought. Once identified, the solutions could prove useful in elaborating evaluation schemes of extracted information in terminographic definition writing.

A theoretical and a methodological framework, backed by empirical data, was presented. This background enabled us to identify two precise questions that should be addressed when designing evaluation schemes for extracted information in the context of terminographic definition writing. (1) The first question was: *What kind of information is relevant and why?* Two hypotheses were put forward, but not examined here: (i) the type of entity defined and (ii) the type of expert domain to which the concept belongs. (2) The second question (*What are the relevance conditions?*) was subdivided into two subquestions: (2a) *How to distinguish between salient (non defining) and potentially relevant (defining) features?* and (2b) *How to select relevant features among potentially relevant ones?* Only the second, (2b), was considered in more detail.

An explanatory proposal concerning feature relevance (question 2b) was put forward, of which one aspect (the extensional dimension) was examined in more detail as an exemplification of the procedure that should be followed to assess the validity and usefulness of the hypothesis for feature evaluation in DE. It was shown that by considering the types of object sets and correlating them with feature coverage, one could, for example, account for different definition structures. Finally, some methodological considerations were discussed to see how proper tests may be devised at a more linguistic level, and thus potentially used to automate the assessment of a feature's relevance.

Applying the techniques used in the example examined here to each attribute-value couple of each dimension and of each feature characteristic, it should be possible to identify the most fruitful hypotheses for relevance determination. Once all the suggested constraints have been examined for their significance in deciding on a feature's relevance, weighting each relevance referential according to what is defined, in which domain, by whom, for whom, in what context and with what purpose could also be considered.

References

[1] G. N. Carlson. *Reference to Kinds in English*. Garland Publishing, New York, 1980.

[2] ISO 704. *Travail terminologique — Principes et méthodes*. ISO, Genève, 2nd edition, 2000.

[3] K. Kageura. *The Dynamics of Terminology: A Descriptive Theory of Term Formation and Terminological Growth*. John Benjamins, Amsterdam, 2002.

[4] S. Laurence and E. Margolis. Concepts and cognitive science. In E. Margolis and S. Laurence, editors, *Concepts : Core Readings*, pages 3–81. The MIT Press, Cambridge Mass. etc., 1999.

[5] S. Löbner, editor. *Understanding semantics*. Understanding language series. Arnold, London, 2002.

[6] É. Machery. *Les concepts ne sont pas une espèce naturelle*. PhD thesis, 2004. [IJN].

[7] R. Martin. *Pour une logique du sens*. Presses Universitaires de France, Paris, 2e édition revue et augmentée edition, 1992.

[8] J. Pearson. *Terms in context*. Studies in corpus linguistics 1. J. Benjamins, Amsterdam etc., 1998.

[9] A. Rey. Synonymie, néonymie et normalisation terminologique. In *Problèmes de la définition et de la synonymie en terminologie*, pages 281–310, Québec, 1982. Girsterm.

[10] A. Rey. *La terminologie : noms et notions*. "Que sais-je ?" n ° 1780, PUF, Paris, 2nd edition, 1992.

[11] A. Rey. *Essays on terminology*. John Benjamins, Amsterdam, Philadelphia, 1995.

[12] H. Rickert. *Science de la culture et science de la nature. Suivi de Théorie de la définition*. Gallimard, Paris, [1915] 1997.

[13] J. Sager and K. Kageura. Concept classes and conceptual structures : Their role and necessity in terminology. *ALFA : Terminology and Special Linguistics*, 7(8):191–216, 1994.

[14] J. C. Sager and M.-C. L'Homme. A model for the definition of concepts : rules for analytical definitions in terminological databases. *Terminology*, 1(2):351–373, 1994.

[15] S. Seppälä. *Composition et formalisation conceptuelles de la définition terminographique*. Mémoire de DEA, ETI, Université de Genève, 2004.

[16] R. Vézina, X. Darras, J. Bédard, and M. Lapointe-Giguère. *La rédaction de définitions terminologiques*. Version abrégée et adaptée par J. Bédard et X. Darras. OQLF, Montréal, 2009.

Linguistic realization of conceptual features in terminographic dictionary definitions

Esperanza Valero
TecnoLeTTra Team
Universitat Jaume I
Castellón, Spain
mvalero@trad.uji.es

Amparo Alcina
TecnoLeTTra Team
Universitat Jaume I
Castellón, Spain
alcina@trad.uji.es

Abstract

In the knowledge society, researchers on lexicography recognize the need to advance in the structure of dictionary information so that it can be understood by people and computers. In relation with dictionary definitions, the adoption of models or templates would be advantageous for the generation of complete definitions and also for the extraction of semantic information from them.

In this study, we manually analyzed specialized dictionary definitions of the ceramic field belonging to the conceptual groups of *ceramic processes* and *ceramic defects* in order to identify the relevant conceptual features, such as physical aspect or function, and the linguistic realization of these features in the definitions.

Results can help to extract information from definitions and to also generate formalized dictionary definitions.

Keywords

Terminographic definition, features, linguistic markers, patterns, conceptual information

1. Introduction

Definitions are a very valuable source of semantic information for different tasks such as the creation of ontologies, lexicography, terminology and natural language processing. Applied research into computational lexicography and terminology is being carried out to recognize definitions in specialized texts and to analyze the semantic relationships that occur in these definitions [19, 20, 14, 17]. Dictionary definitions are other means of obtaining semantic data and of discovering relationships between the concepts of a domain. We can find numerous studies intended to extract information from the definitions in existing dictionaries. Most of them aim to extract taxonomic relationships [5, 11]. However, more effort is needed to exploit the rest of the information present in the definitions, such as specific features and non- hierarchical conceptual relationships [4, 6, 7, 10]. When attempting to extract this information, the biggest problem that automatic systems face is the lack of homogeneity and systematicity in definitions. Most dictionary definitions present inconsistent and incomplete information as well as terms which should be equally treated and which are defined in a very different way [9]. Many authors highlight the need to create more standardized, precise definitions in a format that is understandable for computers and humans alike so that extraction and reusability are easier.

One of the objectives of the ONTODIC[1] project is to develop a system to assist the terminographer in the elaboration of definitions. This system will semi-automatically generate definitions based on a definitional template and a domain ontology. The definitional template will contain the necessary linguistic markers to introduce each feature into the definition. An example of this definitional template for the conceptual group *ceramic tiles* is proposed by Alcina [1]:

"A ceramic tile whose shape is X and size is Y, and is decorated with Z to serve as Q"

Variables x, y, z and q will be replaced by the values that each feature in the concept description acquires.

In this study, our objective is to observe the type of features which are relevant for the description of two conceptual groups (*ceramic processes* and *ceramic defects*) in this domain and to analyze how they are expressed linguistically in the 222 definitions taken from three specialized dictionaries of the ceramic field. We studied the linguistic patterns used in the definitions to introduce features. This set of features and their linguistic realization can, on the one hand, be useful to extract information from dictionaries and, on the other, to generate definitions.

2. Semantic information extraction from texts and dictionary definitions

Several studies have been conducted to automatically identify the related concepts in a corpus. However, most of them focus on the English and French languages with few centering on Spanish [23, 19]. The authors agree that a good way to extract information from texts is by searching for recurrent patterns.

According to [15], *knowledge patterns* are "words, word combinations or paralinguistic features which frequently indicate conceptual relations". The authors distinguish three types of patterns:

1 ONTODIC: Methodology and Technology for the creation of onomasiologic dictionaries based on ontologies. Terminological resources for the e-translation, financed by the Education and Science Ministry (TSI2006-01911).

Lexical patterns which are words that indicate a relation. For example, *is a* can indicate hypernymy.

Grammatical patterns which are combinations of part-of-speech. The NOUN+VERB pattern can indicate the function relation as in the sentence: *L'unité centrale effectue le traitement.*

Paralinguistic patterns which include punctuation, parenthesis, text structure, etc. The authors provide the following example where the questions introduce the hypernymic relation. *Qu'est-ce qu'un réseau ? Un réseau est un ensemble de ressources à la* […]

According to [11], "definitions use a sublanguage of natural language". Dictionary definitions have a more explicit semantic language structure which is accessible for analysis [24].

Conceptually, analytical definitions have two main components: *genus* which reflects the hierarchical conceptual organization in the specialized domain in the definitions and *differentiae* or specific features which distinguish the defined concept from their co-hyponyms and reflect all kinds of relations in a domain.

In [24] it is acknowledged that it is less difficult to extract generic terms from definitions for computational semantics than to extract specific features. It is not possible to find a complete list of the semantic information that can be extracted from definitions in the literature. Sager and L'Homme [18] considered that this part of the definitions is not susceptible to codification following a restricted set of features and that this would be an important step for the systematization of terminographic definitions.

3. Methodology

In this study, we use dictionary definitions to extract the features and relations of concepts. The aim of this study is to discover linguistic patterns that are used in the definitions to denote the specific features of concepts. In order to find regularities among the definitions, we restricted the analysis to the conceptual groups of *ceramic processes* and *ceramic defects* which are described by a limited set of features.

3.1 Definition selection

We analyzed 222 definitions from three dictionaries of the ceramics domain: *Diccionario científico-práctico de la cerámica, Diccionario de cerámica y Terminología de los defectos cerámicos* [12, 13, 21]. These dictionaries are published on paper and have been digitalized [3].

From these dictionaries, we extracted the definitions of the concepts included in the categories of ceramic production processes and defects in the ceramic product. It is important for our study that the conceptual groups to be homogeneous enough to be able to identify a relevant set of features for all of them. In total, we analyzed 135 definitions of the conceptual group *ceramic processes* and 87 *ceramic defects.*

3.2 Identification of conceptual features in the definitions

The definitions in these dictionaries vary considerably as regards to the use of words and their formal structure [2]. Many of them follow the analytical model with the formula Definiendum = genus + differentia, although we can also find descriptions by means of synonyms, paraphrases, etc. In this analysis, we did not focus on the formal aspects of the definitions, but on the conceptual features that they provide. We analyzed the definitions in order to obtain a set of features which are commonly used in the descriptions of the concepts of each conceptual group. We followed the proposal of [16] which distinguishes between the name of the feature and its value. The name of the feature acts as a label that indicates the content of the feature, while the value offers specific information about a concept. For example, cause is the name and friction the value of a feature for the concept abrasion.

This conceptual analysis was carried out by segmenting the information obtained from the definition and by assigning a label or code which describes the type of information that each fragment represents, as seen in Figure 1. In order to carry out this analysis, we used the program for the qualitative analysis named Atlas.ti. This program allows us to: segment the information, assign a descriptive code, create relations between them, obtain graphic representations of the conceptual structure of data and query as in a database.

The result is a list of essential features to describe each category and a set of values for each of these features. The features detected for these categories are as follows (frequency of appearance in the corpus is indicated in the parenthesis):

Production process features are PROCEDURE (69), OBJECTIVE (103), PATIENT (56), MATERIAL STATE (15), INSTRUMENTS (22), PREVIOUS STAGE (6) and NEXT STAGE (8).

Ceramic defect features are PHYSICAL ASPECT (54), ZONE (16), CAUSE (44), PHASE (15), METHOD (5) and PRODUCT (25).

Figure 1. Identifying conceptual features

lágrima: Defecto de aplicación de los esmaltes [STAGE], *consistente en gotones* [PYSICAL ASPECT] *que aparecen en la superficie de la capa* [ZONE] *por haberla depositado incorrectamente, ya sea por usar pincel de punta en vez de pinceleta de punta chata, o por falla en la boquilla del aerógrafo si se esmaltó por pulverización mediante compresor* [CAUSE].

teardrop: defect in the glazing process [STAGE] consisting of drops [PHYSICAL ASPECT] that appear on the surface layer [ZONE] due to incorrect application, either through using a pointed brush instead of a flat brush or because of a fault in the airbrush nozzle if glazing was applied by spray [CAUSE].

hinchamiento: Aumento en las dimensiones o volumen aparente [PHYSICAL ASPECT] *de un articulo* [PRODUCT] *causado por la reacción con el agua o el vapor de agua* [CAUSE] .

bloating: Increase in the dimensions or bulk of the volume [PHYSICAL ASPECT] of an object [PRODUCT] caused by a reaction to water or water vapor [CAUSE].

3.3 Identification of linguistic markers

We marked the linguistic expression that precedes each feature. In some cases it is possible to identify a linguistic marker that introduces a particular feature. As Figure 2 illustrates, the feature OBJECTIVE of a process is introduced into the first definition by the linguistic pattern "que sirve para" (in English "which serves to") and in the second, by the marker "con el fin de" ("for the purpose of").

Figure 2. Examples of linguistic patterns in the definitions

cocción de decoración: cocción que sirve para madurar los efectos decorativos aplicados previamente a las pieza [OBJECTIVE].

decoration firing: firing which serves to mature the decorative effects previously applied to the piece [OBJECTIVE].

compactación de polvos: Operación de condensar los materiales pulverulentos con el fin de obtener productos con la mayor densidad posible [OBJECTIVE].

powder compaction: Operation to condense the powdery materials with the purpose of obtaining products that are as dense as possible [OBJECTIVE].

cocción lenta: cocción que se desarrolla a una velocidad más pequeña de la acostumbrada [PROCEDURE].

slow firing: firing that is done at a lower speed than usual [PROCEDURE].

In many cases, however, it was not possible to identify a linguistic marker to introduce the feature, rather recurrent syntactic structures in the expression of the feature itself. As we can see in Figure 3, the feature PROCEDURE is expressed in both definitions with a sentence in which the main verb is in the gerund.

Figure 3. Recurrent syntactic structures in definitions

colar: Formar una pieza vertiendo un líquido o una masa plástica en un molde [PROCEDURE], *por fraguado o enfriamiento.*

to cast: To form a tile by pouring a liquid or a plastic mass into a mold [PROCEDURE] by setting or cooling.

aclarar, diluir (un color): Rebajar la intensidad de un color añadiéndole componente blanco o agente blanqueante [PROCEDURE].

to lighten , to dilute (a color): To reduce the intensity of a color by adding a white component or a bleaching agent [PROCEDURE].

4. Results

The results obtained in this analysis consist in a set of linguistic markers and recurrent syntactic structures denoting different types of features in the definitions. Linguistic markers introduce the feature; one example would be *en forma de* (Eng. as ……-like) to describe the physical aspect of a defect. Recurrent syntactic patterns are syntactic constructions that mark a particular meaning; for example, the use of gerund constructions to express the mode of carrying out a process.

Table 1: Linguistic markers in the features of *ceramic defects*

Pattern	**Examples from corpora**
PHYSICAL ASPECT	
aparece (2) It appears	aparece rodeada de una envoltura vítrea, acompañada generalmente de una cuerda en forma de cola
en forma de (2) As / …..-like	en forma de motas o agujeritos
*se parece** *a/que parece* (2) It looks like	se parecen a "picaditas de alfiler"
caracterizado por (2) characterized by	caracterizado por una aspereza extrema
presenta (1) It shows	presenta pequeñas arrugas u olas
que se manifiesta (1) which appears as	que se manifiesta en forma de huecos o pequeñas burbujas reventadas

CAUSE	
cuando+oración 8) when+clause	cuando no está bien compensada su fórmula
se deb a* (7) it is due to	se debe a que el coeficiente de dilatación de la pasta es más elevado que el de esmalte
debido a (5) due to	debido a defectos del secado o de la pasta
causad por / porque* (4) caused by / caused because of	causada por fallo mecánico debido a la tensión.
por+ SN (3) because of +NP	por enfriamiento demasiado rápido
por+oración (3) because +clause	por producirse ésta de manera irregular
producida por (2) produced by	producida por un desprendimiento gaseoso anormal
por causa de (1) because of	por causa del enfriamiento demasiado rápido post cocción,
como consecuencia de (1) as a result of	como consecuencia de su disolución parcial
*formad * por* (1) formed by	formada generalmente por la penetración de vidrio entre las partes del molde
originadas por (1) caused by	originadas por impurezas depositadas sobre el vidrio caliente durante su trabajo
procedentes de (1) from	procedente de herrumbre o de otras impurezas
surgida por (1) having emerged by	surgida por retracción del vidrio
da lugar a (1) gives rise to	La presencia de carbón [...] da lugar al “corazón negro”
por efecto de (1) as a result of	por efecto de un recalentamiento excesivo o de la acción de gases
resultado de (1) as a result of	resultado de haber tomado un objeto una forma convexa
consiste en (1) it consists of	Consiste en que una pieza ha perdido fragmentos del vidriado.

STAGE	
durante+SN (7) during + NP	durante el enfriamiento
después de + SN(2) after +NP	después de colocados en la pared
originada (1) caused during	originada durante el prensado
que tiene su origen en(1) which starts in	que tienen su origen en la operación de prensado
en + SN (1) in +NP	en el secado
al+oración (1) when+clause	al levantar el vidrio
METHOD	
cuando+oración (2) when+clause	cuando se esmalta por baño el interior de los jarrones
por +SN (2) by means of +NP	Por vía seca
sufrir (3) suffer	que pueden sufrir los esmaltes
en particular/es (2) particularly	en particular en el vidrio óptico
Presentar (2) present	que suelen presentar los esmaltes cerámicos
*en determined** (1) in certain	En determinadas producciones cerámicas
surgida en (1) having emerged in	surgida en el vidrio
ZONE	
en +SN (8) in +NP	en la superficie
de+SN (3) of+NP	defecto de la superficie
*Aparec** (2) appears	aparece en la superficie de los productos cerámicos
*Situad*a (1) located	Situada en la cara interna
Próxima a (1) near	Próxima a la superficie

Table 2. Recurrent syntactical structures in the features of *ceramic defects*

Feature	Pattern	Examples from corpora
PHYSICAL ASPECT	NP (40)	Velo de burbujas finas
ZONE	Adjective (1)	superficial

Table 3. Linguistic markers in the features of *ceramic processes*

Pattern	Examples from corpora
OBJECTIVE	
para +inf (15) in order to +inf	para dotarle de las propiedades adecuadas.
consistent en* (2) consisting of	consistente en eliminar de las piezas moldeadas las rebabas y otras protuberancias
que sirv para* (1) / it is for	que sirve para madurar los efectos decorativos
PROCEDURE	
*dirigid*a* (1) whose aim is to	Acción dirigida a que un esmalte que da superficies brillantes se transforme…
por+SN (14) by+NP	por inmersión de la pieza en un baño de esmalte
Mediante (3) by means of	Mediante su inmersión en un baño de esmalte
que se logra (1) which is achieved by	Que se logra pasando la mezcla a través del medio adecuado.
*se desarroll** (1) is done	que se desarrolla a una velocidad más pequeña de la acostumbrada.
*se llev*a cabo* (1) is carried out	se lleva a cabo el paso del material a través del tamiz.
PREVIOUS PHASE	
después de (3) after	después de haber sido esmaltada
NEXT PHASE	
previa/o (1) previous	previa homogeneización
antes de (3) before	antes de ser esmaltada
INSTRUMENTS	
mediante (5) by means of	mediante un aerógrafo.
con +SN (2) With +NP	Con piedra o herramienta de acero
en +SN (2) in +NP	en un molde
por +SN (2)/ by+NP	Por prensas mecánicas
empleando (1) using	empleando molinos de rodillos, bolas o guijarros
haciendo uso de using (1)	haciendo uso de un dispositivo (pistola)
*se utiliz** (1) are used	se utilizan resistencias eléctricas
con ayuda de (1) with the help of	con ayuda de un calibre o plantilla
mediante el uso de (1) / by using	mediante el uso de resinas sintéticas de intercambio iónico.
por medio de (1) by means of	por medio de un pistón
se requiere/en (1) is/are required	Se requieren tres tipos de reactivo: espumantes, colectores y controladores
dentro de (1) inside +NP	dentro del molde
PATIENT	
de+ SN (11) of +NP	de unas materias primas
en+SN (6) in + NP	en una pasta o barbotina
a+SN (5) to +NP	al material
se aplica/n a (1) it is applied	se aplica en el esmalte
MATERIAL STATE	
en estado (1) in X state	en estado plástico
en forma de (1) in form of	en forma de barbotina.
en +SN (1) in+NP	en polvo

Table 4. Recurrent syntactical structures in the features of *ceramic processes*

Pattern	Examples from corpora
OBJECTIVE	
infinitive clause (25)	aplicar un color sobre una superficie
PROCEDURE	
simple clause (23)	este se sumerge en la barbotina y se escurre
infinitive (11)	someter un material a la acción del calor a temperaturas altas.
gerund (10)	añadiéndole componente blanco o agente blanqueante.
NP (3)	colocación de piezas unas sobre otras
MATERIAL STATE	
adjective (10)	bizcochado

5. Discussion

We analyzed 222 terminographic definitions belonging to the conceptual groups: *ceramic defects* and *ceramic processes*. In these definitions, we detected 13 types of conceptual features: OBJECTIVE, PROCEDURE, PATIENT, INSTRUMENTS, PREVIOUS PHASE, NEXT PHASE, MATERIAL STATE, PHYSICAL ASPECT, ZONE, CAUSE, PHASE, METHOD and PRODUCT. We identified some patterns which have been used to introduce the features into the definitions. The results of this analysis show problems relating to the variety of the feature expression in the definitions, absence of a common genus in the definitions of the same group and other problems which previous studies identified [15, 8], such as the polysemy of patterns, morphological variants and the non-contiguity of the elements of the pattern. We now go on to explain these problems in the following sections.

- Diversity in the linguistic expression of the features in the definitions

The heterogeneity of these definitions affects both the conceptual and linguistic levels. At the conceptual level, we can see that two *ceramic defects* are not defined when the same features in the definitions are used. For example, in the definition of the defect *dunting*, the features PHYSICAL ASPECT, CAUSE and STAGE are described. However in the definition of the defect *conicity*, only the feature PHYSICAL ASPECT is used to describe this concept. At the linguistic level, we can find a wide range of linguistic markers and structures to express the same kind of feature. For example, the feature PROCEDURE is expressed in 10 definitions of ceramic processes by means of a gerund clause. However, in 23 cases, this feature is a sentence which does not show any common pattern. The only regularity found in these cases is the semantic proximity of the verbs used to describe this feature which are mainly verbs of action (to immerse, to submerge, to project, to pulverize, to attract, etc.), as seen in the following examples: 1. *este se sumerge en la barbotina y se escurre; 2. el esmalte se proyecta sobre la pieza cerámica; 3. el producto bizcochado se sumerge en una suspensión de los ingredientes del esmalte en agua; 4. las partículas que han de ser pulverizadas se les da una carga electrostática opuesta a la de la pieza a esmaltar; esta atrae a las partículas hacia la pieza.* Semantic annotation would be necessary to automatically extract this feature from the definitions.

- Absence of a genus in the definitions

The formal structure of the analyzed definitions does not always respect the established model of genus and differentia. Many definitions do not include the genus or hypernym of the conceptual group. The first position is filled with a specific feature which cannot be preceded by a linguistic marker.

For example, around 70% of the definitions of ceramic defects do not have the genus *defect* or *imperfection*, but include the feature PHYSICAL ASPECT in the first position which is expressed by means of a noun phrase without a linguistic marker, as shown in the following definitions:

> *rebaba. Delgada cresta* (Eng. 'thin crest') [PHYSICAL ASPECT] *en la superficie o en el borde de un objeto, formada generalmente por la penetración de vidrio entre las partes del molde.*
> *desventado. Grieta* (Eng. 'crack') [PHYSICAL ASPECT] *en la pieza cocida por enfriamiento demásiado rápido.*

- Polysemy of linguistic markers

In certain cases, a linguistic marker introduces a different type of feature into the definitions. For example, the structure *por + noun phrase* is used to express the feature CAUSE, as in *por enfriamiento demasiado* (Eng., 'due to fast cooling'), but the PROCEDURE is also used, as in *por inmersión de la pieza* (Eng., 'by immersing the tile'). The same happens with the linguistic marker *consistente en* (Eng. 'consisting in') which precedes the feature PHYSICAL ASPECT of a defect (*consistente en una textura punteada* (Eng. 'consisting in a dotted texture')), as well as the OBJECTIVE of a process (consistente en recubrir el acero o hierro con cinc (Eng. 'consisting in covering the steel or iron with zinc')).

- Morphosyntactic variety of linguistic markers

Many patterns do not show a unique form in the definitions because they can be expressed in any of their morphological variants. This is especially common in the verbal patterns as they appear to be conjugated in different forms, as the example shows.

lo que origina una estructura vesicular en el material que se calienta
puede originarse en el mismo esmalte,
originadas por impurezas,

- Non contiguity of the elements of the marker

We have also found some cases in which the marker is interrupted by another element. For example, the marker *formad* por* appears in the text with an adverb between the verb and the preposition (*formada generalmente por la penetración del vidrio…* (Eng., 'formed generally by the penetration of the glass'))

Results show that the definitions of these dictionaries lack uniformity and that some of the patterns identified in the expression of the features offer some problems that complicate the automatic extraction of information.

6. Conclusions

In this paper we describe the linguistic expression of the different features (such as PHYSICAL ASPECT or CAUSE, included in the definitions of ceramic dictionaries. The aim of the study was to obtain a set of linguistic markers or syntactic patterns for these features. Results reveal a huge heterogeneity in the linguistic realization of the conceptual features in the definitions and the patterns identified show problems such as polysemy and morphological variety.

The ONTODIC project aims to design a tool for computer-assisted terminography which will help the terminographer to create definitions of specialized concepts. These definitions will be based on a template which will guarantee consistency and explicitness. One of the template elements is a restricted set of linguistic markers which will introduce each type of feature into the definitions. This analysis has helped us to observe the naturally occurring markers in the definitions. In future works, we will select one or more linguistic pattern(s) for the expression of each feature in the differentia in our tool for computer-assisted terminography.

7. References

[1] A. Alcina. Metodología y tecnologías para la elaboración de diccionarios terminológicos onomasiológicos, in A. Alcina, E. Valero y E. Rambla (eds.): Terminología y sociedad del conocimiento, Peter Lang, Bern, 2009.

[2] A. Alcina y E. Valero. Análisis de las definiciones del diccionario cerámico científico-práctico. Sugerencias para la elaboración de patrones de definición, *Debate Terminológico*, 4, 2008.

[3] A. Alcina, V. Soler and J. Granell. Translation technology skills acquisition. *Perspectives, Studies in Translatology*, 15 (4), 2007.

[4] H. Alshawi. Analysing the dictionary definitions, in B. Boguraev and T. Briscoe (eds.), 1989.

[5] R.A. Amsler. The structure of the Merriam-Webster pocket dictionary, The University of Texas, 1980.

[6] G. Barnbrook. Defining Language. A local grammar of definition sentences, John Benjamins, Amsterdam, 2002.

[7] C. Barrière. From a Children's First Dictionary to a Lexical Knowledge Base of Conceptual Graphs, Ph.D. Thesis, Ottawa University, 1997

[8] C. Barrière. Investigating the Causal Relation in Informative Texts, *Terminology*, 7(2), 2002.

[9] B. Boguraev and T. Briscoe (eds.). Computational Lexicography for Natural Language Processing. Longman, London, 1989.

[10] N. Calzolari. Acquiring and representing semantic information in a Lexical Knowledge Base. Proceedings of the ACL SIGLEX Workshop on Lexical Semantics and Knowledge Representation, Springer-Verlag, 1992.

[11] N. Calzolari. Detecting patterns in a Lexical Data Base. Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting on Association for Computational Linguistics, Stanford, 1984.

[12] J. F Chiti. Diccionario de cerámica. Buenos Aires, Condorhuasi, 1984.

[13] C. Guillem and M. C. Guillem. Diccionario cerámico científico-práctico, Castellón, Sociedad española de cerámica y vidrio, 1987.

[14] V. Malaisé, P. Zweigenbaum, B. Bachimont. Mining defining contexts to help structuring differential ontologies, *Terminology*, 11 (1), 21–53, 2005.

[15] E. Marshman, T. Morgan and I. Meyer. French patterns for expressing concept relations», *Terminology*, 8 (1), 2002.

[16] I. Meyer, E. Karen and D. Skuce. Systematic concept Analysis within a Knowledge-Based Approach to Terminology, in S. E. Wright and G. Budin (eds.). Handbook of Terminology Management, John Benjamins, Philadelphia, 98-118, 1997.

[17] J. Pearson. Terms in Context. John Benjamins, Amsterdam, 1998.

[18] J. Sager and M.C. L'Homme. A model for the definition of concepts: rules for analytical definitions in terminological databases, *Terminology*, 1(2), 351- 373, 1994.

[19] G. Sierra, R. Alarcón, C. Aguilar and C. Bach. Definitional verbal patterns for semantic relation extraction, *Terminology* 14(1), 74-98, 2008.

[20] G. Sierra and J. McNaught. Design of an onomasiological search system: A concept-oriented tool for terminology, *Terminology*, 6(1), 1-34, 2000

[21] Sociedad española de cerámica y vidrio. Terminología de los defectos del vidrio. Madrid, 1973.

[22] V. Soler and A. Alcina. Patrones léxicos para la extracción de conceptos vinculados por la relación parte-todo en español, *Terminology*, 14(1), 2008.

[23] Y. Wilks, D. Fass, C. Guo, J. McDonald, T. Plate and B. Slator. A tractable machine dictionary as a resource for computational semantics, in B. Boguraev and T. Briscoe (eds.), 1989.

Definition Extraction using Linguistic and Structural Features

Eline Westerhout
Utrecht University
E.N.Westerhout@uu.nl

Abstract

In this paper a combination of linguistic and structural information is used for the extraction of Dutch definitions. The corpus used is a collection of Dutch texts on computing and elearning containing 603 definitions. The extraction process consists of two steps. In the first step a parser using a grammar defined on the basis of the patterns observed in the definitions is applied on the complete corpus. Machine learning is thereafter applied to improve the results obtained with the grammar. The experiments show that using a combination of linguistic (n-grams, type of article, type of noun) and structural information (layout, position) is a promising approach to the definition extraction task.

Keywords

definition extraction, machine learning, grammar, linguistic features, text structure

1 Introduction

Definition extraction is a relevant task in different areas. Most times it is used in the domain of question answering to answer 'What-is'-questions, but it is also used for dictionary building, ontology development and glossary creation. The context in which we apply definition extraction is the automatic creation of glossaries within elearning. Glossaries can play an important role within this domain since they support the learner in decoding the learning object he is confronted with and in understanding the central concepts which are being conveyed in the learning material.

The glossary creation context provides its own requirements to the task. The most relevant one is constituted by the corpus of learning objects which includes a variety of text genres (such as manuals, scientific texts, descriptive documents) and also a variety of writing styles that pose a real challenge to computational techniques for automatic identification and extraction of definitions together with the headwords. Our texts are not as structured as those employed for the extraction of definitions in question-answering tasks which most times include encyclopedias and Wikipedia. Furthermore, some of our learning objects are relatively small in size, thus our approach has not only to favor precision but also recall. That is, we want to make sure that as many as possible definitions present in a text are proposed to the user for the creation of the relevant glossary. Therefore, the extraction of definitions cannot be limited to sentences consisting of a subject, a copular verb and a predicative phrase, as is often the case in question-answering tasks, but a much richer typology of patterns needs to be identified than in current research on definition extraction.

Different approaches for the extraction of definitions can be distinguished. We use a sequential combination of a rule-based approach and machine learning to extract them. As a first step a grammar is used to match sentences with a definition pattern and thereafter, machine learning techniques are applied to filter out those sentences that – although they have a definition pattern – do not qualify as definitions.

Our work has several innovative aspects compared to other work in this area. First, we address less common definition types in addition to 'to be' definitions. Second, we apply a machine learning algorithm designed specifically to deal with imbalanced datasets, which seems to be more appropriate for us because we have data sets in which the proportion of 'yes'-cases is extremely low. The third innovative aspect on which this paper focuses has to do with the combination of different types of information for the extraction of definitions. Not only linguistic information (n-grams, type of article, type of noun) has been used, but also experiments with structural and textual information have been carried out (position, layout).

The paper is organized as follows. Section 2 introduces relevant work in definition extraction, focusing on the work done within the glossary creation context. Section 3 describes the data used in the experiments and the definition categories we distinguish. In section 4 the way in which grammars have been applied to extract definitions and the results obtained with them are discussed. Section 5 talks about the machine learning approach, covering issues such as the classifier, the features and the experiments. Section 6 reports and discusses the results obtained in the experiments. Section 7 provides conclusions and presents some future work.

2 Related research

Research on definition extraction has been pursued mainly in the context of automatic dictionary building from text, question-answering and ontology development. Initially, mainly pattern-based methods were used to extract definitions (cf. [12, 15, 16, 19]) but recently, several researchers have started to apply also machine learning techniques and combinations of

Workshop On Definition Extraction 2009 - Borovets, Bulgaria, pages 61–67,

pattern-based methods and machine learning in this area (cf. [2, 9, 11]). [20] provides an overview of the work done in the different areas and compares it to the task within the glossary creation context.

Definition detection approaches developed in the context of question-answering tasks are often definiendum-centered, that is, they search for definitions containing a given term. Our approach, in contrast, is connector-centered, which means that we search for verbs or phrases that typically appear in definitions with the aim of finding the complete list of all definitions in a corpus independently of the defined terms. Despite the challenges that the eLearning application involves, we believe that the techniques for the extraction of definitions developed within the Natural Language Processing and the Information Extraction communities can be adapted and extended for our purposes.

Our work on definition extraction started within the European LT4eL project. Within the scope of this project experiments for different languages have been carried out. [13] describe experiments on definition extraction in Slavic languages and present the results obtained with Bulgarian, Czech and Polish grammars. The three grammars show varying degrees of sophistication. The more sophisticated the grammar, the more patterns are covered. Although the recall improves when more rules are added, the precision does not drop and is comparable for the three languages (22.3-22.5%).

For Polish, [10, 14, 7] put efforts in outperforming the pattern-based approach using machine learning techniques. To this end, [10] describe an approach in which the Balanced Random Forest classifier is used to extract definitions from Polish texts. They compare the results obtained with this approach to results obtained with experiments on the same data in which grammars were used [14] and to results of experiments with standard classifiers [7]. The best results are obtained with the approach designed for dealing with imbalanced datasets. The differences with my approach are that (1) they used either only machine learning or only a grammar and not a combination of the two, (2) they did not distinguish different definition types and (3) they only used relatively simple features, such as n-grams.

[3] applies Genetic Algorithms to the extraction of English 'to be' definitions. Her experiments focus on assigning weights to a set of features for the identification of such definitions. These weights act as a ranking mechanism for the classification of sentences, providing a level of certainty as to whether a sentence is actually a definition or a non-definition. They obtain a precision of 62% and a recall of 52 % on the extraction of is definitions by using a set of features such as 'has keyword' and 'contains 'is a'.

[8] focus on the extraction of Portuguese 'to be' definitions. First, a simple grammar is used to extract all sentences in which the verb 'to be' is used as main verb. Because their corpus is heavily imbalanced and only 10 percent of the sentences are defintions, they investigate which sampling technique gives the best results and present results from experiments that seek to obtain optimal solutions for this problem.

Previous experiments for Dutch focused on using a grammar [22], and using several combinations of machine learning and a grammar to extract definitions [21, 23, 20]. A comparison of a standard classifier (naive Bayes) and the Balanced Random Forest (BRF) classifier showed that, especially for the more imbalanced data sets, the BRF classifier outperforms the naive Bayes classifier [20]. In all these previous experiments the features used were either only n-grams or a combination of n-grams and linguistic features.

3 Data

Definitions are expected to contain at least three parts. The definiendum is the element that is defined (Latin: that which is to be defined). The definiens provides the meaning of the definiendum (Latin: that which is doing the defining). Definiendum and definiens are connected by a verb or punctuation mark, the connector, which indicates the relation between definiendum and definiens [19].

Based on the connectors used in the 603 manually annotated patterns, four common definition types were distinguished. The first type are the definitions in which a form of the verb 'to be' is used as connector (called 'is definitions'). The second group consists of definitions in which a verb (or verbal phrase) other than 'to be' is used as connector (e.g. to mean, to comprise). It also happens that a punctuation character is used as connector (most times the colon), such patterns are contained in the third type. The fourth category contains the definitory contexts in which relative or demonstrative pronouns are used to point back to a defined term that is mentioned in a preceding sentence. The definition of the term then follows after the pronoun. Table 1 shows an example for each of the four types.

4 Grammar

The first part of the extraction process is rule-based in our approach. Based on the part-of-speech tag patterns observed in the development part of the corpus a grammar was written to detect the four types of definitions. For a proper extraction of both sentences of multi-sentence pronoun definitions, anaphora resolution would have to be included in the system. As this is a completely different topic, we decided to restrict ourselves to only looking at the part of the definition containing the pronoun and connector verb (phrase). When the tool is integrated into the Learning Management System, it shows for each definition candidate one sentence to the left and one sentence to the right to see the context in which it is used. For the multi-sentence pronoun definitions this makes it possible to see which term is defined in the previous sentence and to select it manually.

The XML transducer LXTransduce developed by [18] has been used to match the grammars against files in XML format. LXTransduce is an XML transducer that supplies a format for the development of grammars which are matched against either pure text or XML documents. The grammars are represented in XML using the lxtransduce.dtd DTD, which is part

Type	Example sentence
is	Gnuplot is een programma om grafieken te maken *'Gnuplot is a program for drawing graphs'*
verb	E-learning omvat hulpmiddelen en toepassingen die via het internet beschikbaar zijn en creatieve mogelijkheden bieden om de leerervaring te verbeteren . *'eLearning comprises resources and application that are available via the Internet and provide creative possibilities to improve the learning experience'*
punctuation	Passen: plastic kaarten voorzien van een magnetische strip, die door een gleuf gehaald worden, waardoor de gebruiker zich kan identificeren en toegang krijgt tot bepaalde faciliteiten. *'Passes: plastic cards equipped with a magnetic strip, that can be swiped through a card reader, by means of which the identity of the user can be verified and the user gets access to certain facilities. '*
pronoun	Dedicated readers. Dit zijn speciale apparaten, ontwikkeld met het exclusieve doel e-boeken te kunnen lezen. *'Dedicated readers. These are special devices, developed with the exclusive goal to make it possible to read e-books.'*

Table 1: *Examples for each of the definition types*

of the software. A sentence is classified as a definition sentence if the parsing algorithm finds a match in this sentence of at least one token (not necessarily spanning the whole sentence).

type	R	P	F	F_2
is	0.83	0.36	0.50	0.58
verb	0.75	0.45	0.56	0.61
punctuation	0.93	0.07	0.13	0.18
pronoun	0.64	0.09	0.16	0.21
all	0.79	0.16	0.27	0.34

Table 2: *Results with the grammar*

Table 4 shows the results obtained with the grammar. As can be seen from this table, the precision is quite low for all types, especially for the punctuation and pronoun types. The grammar rules were thus not specific enough to filter the incorrect sentences. To improve these low precision scores, machine learning has been applied on the grammar results.

5 Machine learning

The datasets obtained with the grammar are imbalanced, especially for the punctuation and pronoun definitions. Our interest leans towards correct classification of the smaller class (the 'positive' class), that is, the class containing the definitions. Therefore, a classifier specifically designed to deal with imbalanced datasets has been used, namely the Balanced Random Forest classifier. After describing how this classifier works, the features and feature settings are set out.

5.1 Balanced Random Forest Classifier

The Random Forest classifier is a decision tree algorithm, which aims at finding a tree that best fits the training data. Whereas normally the underlying tree is a CART tree, in the Weka package it is a modified variant of REPTree. The Weka algorithm follows the same methods of introducing randomness and voting of models. At the root node of the tree the feature that best divides the training data is used. In the Random Forest classifier [5] the *Gini index* is used as splitting measure.

In the Random Forest classifier there is not just one tree used for classification but an ensemble of trees [4]. The 'forest' is created by using bootstrap samples of the training data and random feature selection in tree induction. Prediction is made by aggregating the predictions of the ensemble. This idea behind Random Forest can be used in other classifiers as well and is called *bagging* (**b**ootstrap **agg**regat**ing**).

A disadvantage of the Random Forest approach is that when data are extremely imbalanced, there is a significant probability that a bootstrap sample contains few or even none of the minority class. As a consequence, the resulting tree will perform poor when predicting the minority class. To solve this problem, [6] proposed the Balanced Random Forest classifier. This is a modification of the Random Forest method specifically designed to deal with imbalanced data sets using down-sampling. In this method a adapted version of the bagging procedure is used, the difference being that trees are induced from *balanced* down-sampled data. The procedure of the Balanced Random Forest (BRF) algorithm is described by [6]:

1. For each iteration in random forest, draw a bootstrap sample from the minority class. Randomly draw the same number of cases, with replacement, from the majority class.
2. Induce a classification tree from the data to maximum size, without pruning. The tree is induced with the CART (Classification and Regression Trees) algorithm [5], with the following modification: at each node, instead of searching through all variables for the optimal split, only search through a set of m randomly selected variables[1].
3. Repeat the two steps above for the number of times desired. Aggregate the predictions of the ensemble and make the final prediction.

5.2 Features

The features that have been used can be divided into five categories. Several combinations of these features resulted in 16 settings.

[1] [4] experimentend with $m = 1$ and a higher value of m and concluded that the procedure is not very sensitive to the value of m. The average absolute difference between the error rate using F=1 and the higher value of F is less than 1%

1. **Text properties**: these include various types of n-grams with different values for n.

2. **Syntactic properties**: features of this category give information on syntactic properties of the sentences, in these experiments the type of article used in definiens and definiendum are considered.

3. **Word properties**: in this category information on specific words is included, in these experiments, whether the noun in the definiens is a proper or a common noun.

4. **Position properties**: these include several features which give information on the place in the document where the definition is used.

5. **Lay-out properties**: this category contains features on layout information used in definitions.

N-grams

In many text classification tasks n-grams are used for predicting the correct class (cf. [1] and [17]). For the classification of definitions two types of n-grams have been used, with n being 1, 2 or 3. We used Part-of-Speech tag (PoS-tag) n-grams. The tagger used distinguished 9 parts of speech: adjective, adverb, article, conjunction, interjection, noun, numeral, preposition, pronoun, verb. In addition it used the tag 'Misc' for unknown words and 'Punc' for punctuation marks.

Articles

[9] investigated whether there is a connection between the type of article used in the definiendum (definite, indefinite, other) and the class of sentences (definition or non-definition). Although our definition corpus contains less structured texts than the data used by [9] (Wikipedia texts), part of the figures are quite similar for our data (table 3). In the Wikipedia sentences, the majority of subjects in definition sentences did not have an article (63%), which is the same in our corpus (62%). A difference with their data is the proportion of indefinite articles, which is 25% in our data and 13% in the data from [9].

	definition	**non-definition**
definite	12.8%	44.4%
indefinite	25.0%	8.3%
no article	62.2%	43.7%
other	0%	3.6%
	100%	100%

Table 3: *Proportions of article types used in definiendum of* is-*definitions*

The differences in distribution observed for the *is*-definitions is not seen to the same extent for the verb and punctuation definitions. In the verb definition candidates, for instance, both in definitions and non-definitions, definite articles tend to be used. However, also for these types there is a difference between definitions and non-definitions with respect to this feature.

The article used in the predicate complement has also been included. Again, we observe similarities and differences between our data and the data from [9]. In both data sets the vast majority of articles tends to be indefinite at the start of the definiens (72% and 64%), which is quite different from the proportions for the non-definitions (30% and 29%). Differences between the two data sets are the proportion of definite articles in the definitions group (15% and 23%) and the proportion of no articles in the non-definitions (18% and 1%), which is much higher in the LT4eL data set.

	definitions	**non-definitions**
definite	14.7%	30.0%
indefinite	71.8%	30.0%
no article	9.0%	18.7%
other	4.5%	21.3%
	100%	100%

Table 4: *Proportions of article types used at start of definiens in* is-*definitions*

Nouns

Nouns can be divided into two types, namely proper nouns and common nouns. Unfortunately, with our linguistic annotation tools it was not possible to get more detailed information about the type of proper noun (e.g. person, location), so we can only distinguish between proper and common nouns.The distribution of these types is different for definitions and non-definitions, especially for *is*-definitions. In the *is*-definitions the proportion of proper nouns in the definiendum is considerably higher for the definitions than for the non-definitions (53% versus 31%). For the other definition types the difference observed is much smaller.

Layout

Because definitions contain important information you might expect special layout features (e.g. bold, italics, underlined) to occur more often in definitions than in non-definitions. Because in our data information on the original layout of the documents has been stored per word it was possible to check whether this was the case. No other research on definitions included this property in their research as far as we know. For each of the sentences it was indicated whether a specific layout feature was used in the definiendum. Because of the small numbers for some of the properties we decided to combine all layout features into one group. A comparison shows that *is*, *verb* and *punctuation* definition sentences contain significantly more layout information in the definiendum than non-definition sentences.[2] For each of the definition types the proportion of layout information is about twice as high in definitions than in non-definitions.

Position

[9] in their reseach on definition extraction from Wikipedia texts reduced the set of definition candi-

[2] The pronoun definitions were not included in this investigation, because the definiendum of these sentences is often not in the same sentence as the definiens

dates extracted with the grammar by selecting only the first sentences of each document as possible candidates. It seems that Google's define query feature also relies heavily on this feature to answer definition queries. However, as [9] also state, the first position sentence is likely to be a weaker predictor of definition versus non-definition sentences for documents from other sources, which are not as structured as Wikipedia. The texts from the LT4eL corpus are such less structured texts and therefore using this restriction would not be a good decision when dealing with these documents. In addition to being less structured, they are also often longer and contain on average 10.6 definitions, so applying the first sentence restriction would cause a dramatic decrease of recall and make it impossible to fulfil our aim of extracting as much definitions as possible because at most one sentence per document would be extracted using this method.

Although we thus cannot use the same restriction, it is nevertheless possible to include information on the position of the definition candidate in a document as feature in the machine learning experiments to see whether it helps the classifier in predicting the correct class. To this end, three types of positional information were included in the features, namely information on the position of the sentence within the paragraph, information on the position of the definition within the sentence and information on the (relative and absolute) position of the definiendum compared to other occurrences of the term in the document.

Position in paragraph Each document is divided into paragraphs which are again divided into sentences. It is thus possible to see where in the paragraph a definition is used. When we consider each paragraph as a separate block of information, we would expect definitions to appear at the beginning of such a block. The fact that sentence position is such a strong predictor in Wikipedia articles supports this idea.

The first property related to position in paragraph is the absolute position of the definition sentence within the paragraph. When we compare definitions and non-definitions with respect to this feature we see that for three of the four definition types the absolute position is lower for the definitions. Only of the pronoun definitions there is no significant difference. The pronoun definitions tend to be used later on in the paragraph compared to the non-definitions for this type. This might be caused by the fact that they are used more often at the second position of the paragraph where the term is mentioned in the first sentence.

In addition to the absolute position of a sentence, we also included a score on the relative position taking into account the number of sentences in a paragraph, because the beginning of a paragraph is a relative property. When we compare the scores on this property for definitions and non-definitions, for three of the four types there is a significant difference, only the result for the *punctuation*-definitions is not significant.

Position in sentence When we look at the four definition types, one of the differences observed is the place in the sentence where it can start and end. Whereas *is* and *verb* definitions tend to span a complete sentence, the rules for punctuation definition are less strict for this feature. On the basis of this observation I investigated whether information on this could be used to distinguish definitions from non-definitions.

In addition to this, a second reason has to do with the conversion from original document to XML document. During this process sentences were split automatically and marked as `<s>`. However, not all sentences were splitted correctly, because the sentence splitter tool made errors sometimes which were not corrected manually. Therefore, an extra rule had to be used to detect the beginning of a sentences saying that each word starting with a capital could indicate the start of a sentence.

The position is given by indicating the number of tokens in the `<s>` before the definition starts. For all definition types, the absolute position of the definition candidate within the sentence is significantly lower for definitions than for non-definitions.

Position of definiendum When a term is defined, one would expect that it has not been used a lot of times before it is explained in the definition. Although it is possible that it has been used two or three times before already (e.g. in title of document, table of contents or heading), intuitively you would expect it to be used more after it has been explained. Based on this intuition three measures have been included.

The first two are the absolute number of occurrences of the term before and after it is used in the definition candidate. For all types the average number of occurrences before is lower for definitions. This difference is significant for all types except for the *is*-definitions. The number of occurrences of the term after it has been defined seems to be a less good predictor and is only significantly lower for the *is*-definitions. When we look at the relative position of the definiendum the score is significantly lower for the definition sentences for all types except the *is*-definitions for which there is no difference observed.

5.3 Feature settings

The first setting are the n-grams of part-of-speech tags. This setting is the baseline to which all other settings are compared. The four types of features – articles, nouns, position and layout – have been combined in all possible ways resulting in 16 settings in total. In the second group the four types of feature settings were tried separately (setting 2 to 5). Settings 6 to 11 are all possible combinations of two of the four settings. Then there are four settings (12 to 15) in each of which three types were combined and in the last setting all four types are integrated. Table 5 shows the settings.

6 Results

The final results after applying both the grammar and machine learning are shown in table 6. The sentences not detected with the grammar rules could of course not be retrieved anymore, and as a consequence the recall after applying machine learning is always lower

	IS				VERB				PUNCTUATION				PRONOUN			
setting	R	P	F	A	R	P	F	A	R	P	F	A	R	P	F	A
1.	0.57	0.49	0.53	0.60	0.58	0.54	0.56	0.56	0.51	0.16	0.24	0.74	0.40	0.15	0.22	0.64
2.	0.74	0.56	0.64	0.66	0.49	0.53	0.51	0.54	0.50	0.13	0.21	0.70	0.55	0.17	0.26	0.61
3.	0.49	0.47	0.48	0.58	0.49	0.43	0.46	0.43	0.43	0.11	0.18	0.68	0.49	0.21	0.29	0.70
4.	0.57	0.50	0.54	0.61	0.52	0.53	0.53	0.54	0.47	0.13	0.20	0.70	0.47	0.19	0.27	0.67
5.	0.17	0.52	0.26	0.61	0.15	0.56	0.24	0.53	0.39	0.14	0.21	0.76	0.57	0.09	0.15	0.21
6.	0.70	0.56	0.62	0.66	0.49	0.61	0.55	0.60	0.60	0.11	0.19	0.58	0.56	0.18	0.27	0.62
7.	0.64	0.63	0.64	0.71	0.56	0.56	0.56	0.58	0.53	0.15	0.24	0.73	0.47	0.22	0.30	0.72
8.	0.74	0.57	0.64	0.68	0.44	0.54	0.49	0.54	0.52	0.13	0.21	0.68	0.53	0.17	0.26	0.62
9.	0.54	0.52	0.53	0.62	0.56	0.56	0.56	0.57	0.50	0.15	0.23	0.73	0.45	0.18	0.26	0.68
10.	0.53	0.47	0.50	0.58	0.22	0.46	0.30	0.49	0.42	0.16	0.24	0.78	0.53	0.20	0.29	0.67
11.	0.57	0.52	0.54	0.62	0.52	0.51	0.51	0.52	0.48	0.14	0.21	0.72	0.44	0.19	0.26	0.68
12.	0.63	0.62	0.62	0.70	0.56	0.58	0.57	0.59	0.53	0.17	0.26	0.75	0.47	0.23	0.31	0.73
13.	0.69	0.57	0.62	0.67	0.51	0.64	0.57	0.62	0.53	0.14	0.22	0.70	0.57	0.19	0.29	0.64
14.	0.66	0.64	0.65	0.72	0.59	0.57	0.58	0.58	0.54	0.16	0.24	0.73	0.46	0.22	0.30	0.72
15.	0.57	0.53	0.54	0.63	0.53	0.52	0.52	0.53	0.45	0.14	0.21	0.73	0.47	0.20	0.28	0.70
16.	0.63	0.63	0.63	0.71	0.56	0.57	0.56	0.58	0.47	0.15	0.23	0.75	0.42	0.22	0.29	0.73

Table 6: *Final results after applying grammar and machine learning*

#	setting
1.	n-grams
2.	article
3.	noun
4.	position
5.	layout
6.	article + noun
7.	article + position
8.	article + layout
9.	noun + position
10.	noun + layout
11.	position + layout
12.	article + noun + position
13.	article + noun + layout
14.	article + position + layout
15.	noun + position + layout
16.	article + noun + position + layout

Table 5: *The sixteen feature settings*

than the recall obtained in the first step. For each experiment four measures are reported. The first three are the recall, precision, and f-score of the definition class. The fourth score is the overall classification accuracy. The separate results for the non-definition class are not shown. As the aim of the experiments is to improve the precision obtained with the grammar, this is the most important measure. However, recall and accuracy may not become too low and therefore also recall, f-score and accuracy are reported.

For each of the types it is described in this section how the results should be interpreted and to which extent the settings can compete with setting 1 (n-grams).

6.1 Results per type

Is definitions The first block of information in table 6 shows the results for the is definitions. We see that for this type the article is the best feature for classification. Using only this feature gives better results than the results obtained with the n-grams. The second best individual feature is the information on position, although for this type the results with the n-grams are almost the same. A combination of article, noun and position (setting 14) gives the best result, which is equally good as the result obtained with a combination of article and position (setting 7) and a combination of all feature settings (setting 16).

For the layout setting the recall is very low, which is not strange given the fact that only in a small subset of the definitions there was special layout used. Although there is a slight improvement when it is used in combination with other features, the added value is not big. Adding the noun to other settings generally leads to either lower or similar classification results.

The maximum improvement of precision compared to the precision obtained with the grammar is 77.8% (setting 14).

Verb definitions The second group of definitions in table 6 are the verb definitions. For this type none of the individual settings outperforms the baseline set by the n-grams. The best feature here is position. Using a combination of features makes it possible to perform better than the n-grams. The highest precision is obtained with setting 13, which is a combination of article, noun and layout. The results with the layout setting are comparable to the results for the is definitions. The grammar precision for this type was 0.45 so the maximum improvement is 42.2% (setting 13).

Punctuation definitions For the punctuation definitions the accuracy is highly determined by the non-definitions, as these constitute over 90% of the data set. For the individual feature settings the best precision and accuracy are obtained with the layout setting, however, the recall is quite low for this type. Only one of the settings gives better results than the n-grams, namely setting 12 (article, noun and position). The maximum improvement of precision compared to the precision obtained with the grammar is 142.9% with this setting.

Pronoun definitions Just as for the punctuation definitions, the pronoun definitions data set is highly imbalanced. The noun is the most important individual feature setting, which is surprising as many of these definitions do not have a definiendum. In most settings the recall improves compared to the result on this score of the n-grams, but it often goes with a drop of precision. An overall improvement compared to the base line is observed in most of the settings, especially

in setting 7 (article and position) and 10 (noun and layout) and the best result is obtained with setting 12 (article, noun and position), which is considerably better than the result of setting 1. With this setting the increase of the precision score compared to the precision obtained with the grammar is 155.6% (setting 12).

6.2 General observations

When looking from the perspective of the settings, we see that the article and position in general are the best features. The problem with the layout feature setting mainly is that the recall obtained with it is quite low. Also, adding it as an extra feature to other settings does not lead to much improvement of these results.

A second general observation is that for none of the types the best results are obtained when a combination of all features is used. It is thus not the case that the more information is included the better results will be obtained. For all types one or more feature settings outperform the n-grams results.

7 Conclusions and future work

The influence of the inclusion of linguistic and structural features on classification accuracy differs per type and per combination of settings. Except for the layout setting all individual settings perform well on at least one of the definition types. Combining the different feature settings generally improves the results.

The precision improved in all cases. The two types on which the grammar performed best (is and verb) showed a substantial improvement of 77.8% and 44.2%. And even though precision was still low for punctuation and pronoun patterns after applying machine learning, the percentual improvement was huge for these types (142.9% and 155.6% respectively).

The fact that it is possible to obtain better results with linguistic and structural features than with part-of-speech n-grams is encouraging for several reasons. First, because it shows that it makes sense to use other information in addition to linguistic information (position and lay-out settings) and to structure the linguistic information (article and noun settings). A second issue is that those features provide us more insight on how definitions are used, which is relevant for research on definitions.

As the results are promising, future work will proceed in this direction. We plan to conduct experiments in which other feature settings that go beyond use of linguistic information are used in addition to the settings discussed in this paper. An example of such a setting is the importance of words in a text ('keywordiness'). Another future experiment will investigate whether the number of included n-grams (in these experiments we included all n-grams) can be decreased to lower the computational load while keeping the same results. Initial experiments with 100 n-grams for the is definitions did not show much decrease in performance.

References

[1] R. Bekkerman and J. Allan. Using bigrams in text categorization. Technical report, Technical Report IR, 2003.

[2] S. Blair-Goldensohn, K. R. McKeown, and A. Hazen Schlaikjer. *New Directions In Question Answering*, chapter Answering Definitional Questions: A Hybrid Approach. AAAI Press, 2004.

[3] C. Borg. *Automatic definition extraction using evolutionary algorithms*. PhD thesis, University of Malta, 2009.

[4] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[5] L. Breiman, J. Friedman, R. Olshen, C. Stone, L. Breiman, J. Freidman, R. Olshen, and C. Stone. *Classification and regression trees*. Wadsworth, 1984.

[6] C. Chen, A. Liaw, and L. Breiman. Using Random Forest to learn imbalanced data. Technical Report 666, University of California, Berkeley, 2004.

[7] L. Degórski, M. Marcińczuk, and A. Przepiórkowski. Definition extraction using a sequential combination of baseline grammars and machine learning classifiers. In *Proceedings of LREC 2008*, 2008.

[8] R. Del Gaudio and A. Branco. Extraction of definitions in portuguese: An imbalanced data set problem. In *Proceedings of Text Mining and Applications at EPIA 2009*, 2009.

[9] I. Fahmi and G. Bouma. Learning to identify definitions using syntactic features. In R. Basili and A. Moschitti, editors, *Proceedings of the EACL workshop on Learning Structured Information in Natural Language Applications*, 2006.

[10] L. Kobyliński and A. Przepiórkowski. Definition extraction with balanced random forests. In B. Nordström and A. Ranta, editors, *Advances in Natural Language Processing: Proceedings of the 6th International Conference on Natural Language Processing, GoTAL 2008*, pages 237–247. Springer Verlag, LNAI series 5221, 2008.

[11] S. Miliaraki and I. Androutsopoulos. Learning to identify single-snippet answers to definition questions. In *Proceedings of COLING 2004*, pages 1360–1366, 2004.

[12] S. Muresan and J. Klavans. A method for automatically building and evaluating dictionary resources. In *Proceedings of the Language Resources and Evaluation Conference*, 2002.

[13] A. Przepiórkowski, L. Degórski, M. Spousta, K. Simov, P. Osenova, L. Lemnitzer, V. Kubon, and B. Wójtowicz. Towards the automatic extraction of denitions in Slavic. In *Proceedings of BSNLP workshop at ACL*, 2007.

[14] A. Przepiórkowski, M. Marcińczuk, and L. Degórski. Dealing with small, noisy and imbalanced data: Machine learning or manual grammars? In *Proceedings of TSD 2008*, 2008.

[15] H. Saggion. Identifying definitions in text collections for question answering. In *Proceedings of the Language Resources and Evaluation Conference*, 2004.

[16] A. Storrer and S. Wellinghof. Automated detection and annotation of term definitions in German text corpora. In *Proceedings of LREC 2006*, 2006.

[17] C. Tan, Y. Wang, and C. Lee. The use of bigrams to enhance text categorization. *Information Processing and Management*, 38(4):529–546, 2002.

[18] R. Tobin. Lxtransduce, a replacement for fsgmatch, 2005. `http://www.ltg.ed.ac.uk/ richard/ltxml2/lxtransduce-manual.html`.

[19] S. Walter and M. Pinkal. Automatic extraction of definitions from German court decisions. In *Proceedings of the workshop on information extraction beyond the document*, pages 20–28, 2006.

[20] E. Westerhout. Extraction of definitions using grammar-enhanced machine learning. In *Proceedings of the Student Research Workshop at EACL 2009*, pages 88–96, Athens, Greece, 2009. Association for Computational Linguistics.

[21] E. Westerhout and P. Monachesi. Combining pattern-based and machine learning methods to detect denitions for elearning purposes. In *Proceedings of RANLP 2007 Workshop "Natural Language Processing and Knowledge Representation for eLearning Environments"*, 2007.

[22] E. Westerhout and P. Monachesi. Extraction of Dutch definitory contexts for elearning purposes. In *Proceedings of CLIN 2006*, 2007.

[23] E. Westerhout and P. Monachesi. Creating glossaries using pattern-based and machine learning techniques. In *Proceedings of LREC 2008*, 2008.

Author Index