English summary

The book presents the summary of corpus-based research on coreference resolution for Polish. Subsequent phases of the work are described in a project-based manner, reflected in the order of chapters: after outlining the initial assumptions (chapter 1) and the state of current theoretical and practical knowledge in the field (chapter 2), the model of general reference is constructed (chapter 3) and used for annotation of the corpus of reference relations (chapter 4). The corpus provides the data for development and evaluation of automatic reference detection tools (chapters 5 and 6). The concluding chapters put the research in a broader context of discourse modelling (chapter 7) and provide a summary of the results and development perspectives.

1 Introduction

The introductory chapter brings the basic definitions of reference, anaphora, coreference and bridging, establishing the Polish terminology of the field. The motivation for the work is presented, situating the phenomenon of linguistic reference at the heart of semantic text structure and pointing out that despite numerous existing theoretical works, computational processing of reference in Polish was not systematically carried out until the recent years. Only the latest advances of corpus linguistics in Poland and the development of numerous syntactic and semantic natural language processing tools (morphological analysers, named entity recognizers, dependency parsers, etc.) made it possible to verify existing hypotheses on a much larger scale and create tools for end-to-end coreference resolution.

The scope of the work is limited to decoding nominal coreference, i.e. clustering textual fragments (*mentions*) expressed with nominal constructs (phrases, pronouns, named entities, elided subjects) into equivalence classes based on their reference to different discourse-world entities. Since coreferent mentions can be expressed with a variety of linguistic means (such as synonyms, hypernyms, neologisms, proper names or even idiolectal expressions), this task is considered one of the most difficult in natural language processing.

The methodology of this work is based on a corpus-based approach which requires establishing a formal model of reference relations, carrying out annotation of language data following this representation and using the data set for development, training and formal evaluation of implemented tools.

2 Related work

Chapter 2 starts with a short synthesis of findings from the Polish theoretical literature influencing our computational-linguistic work. The broad concept of *referencing elements* being instantiated in texts not only as single phrases, but also as larger pieces of text (whole sentences or even paragraphs) has been borrowed from Klemensiewicz (1937). The basic categorization of the reference expressions, implemented as nouns, nominal groups, proper names or pronouns was inspired by Bellert (1971).

Portions of the two most extensive taxonomies of reference relations presented by Topolińska (1984) and Paduczewa (1992) were adopted as the starting point for the final classification. The line between referencing and non-referencing expressions was drawn following the approach of Langacker (2008), Vater (2009) and Kunz (2010), assuming the prevalence of the discourse world over the real world in decoding reference, i.e. treating all nominal phrases as potentially referential. A description of indirect relations combining several existing classifications was also presented, starting from the best known one by Clark (1977) and several others, summarized in a paper by Gardent et al. (2003).

Lexical features of referential relations were inspired by the work of Pisarkowa (1969), Fontański (1986) and Grzegorczykowa (1996). Numerical features reflecting word order and inter-sentential position of mentions were taken from Szwedek (1975), Honowska (1984) and Duszak (1986). Following Gajda (1990), we analyzed the density of reference expressions depending on the text genre and similarly to Dobrzyńska (1996) we investigated the variation of stylistic quality of text (here measured by its readability).

The chapter also contains an extensive list of the largest foreign corpora annotated with referential relations and existing Polish reference-related corpora (Filak 2006, Marciniak 2010, Broda et al. 2012b). Previous attempts of anaphora and coreference resolution for Polish (Mitkov and Styś 1997, Kulików et al. 2004, Abramowicz et al. 2006, Filak 2006, Broda et al. 2012b, Kaczmarek and Marcińczuk 2017) are

investigated together with our own translation- and projection-based experiment (Ogrodniczuk 2013).

Different variants of foreign computer-based implementations of reference influencing our target solution are presented in order to outline the history of linguistic engineering work to date. Starting with early attempts using syntactic rules (Hobbs 1976), centering theory (Grosz 1977, Sidner 1979, Brennan et al. 1987) or knowledge poor approaches (Mitkov and Styś 1997), we shortly present supervised machine learning algorithms (see, e.g., Connolly et al. 1994, McCarthy and Lehnert 1995, Kehler 1997, Soon et al. 1999, 2001, Ng and Cardie 2002, Rahman and Ng 2009), high-precision sieve-based methods (see, e.g., Haghighi and Klein 2009, Raghunathan et al. 2010, Lee et al. 2011), hybrid approaches (Denis and Baldridge 2008, Chen and Ng 2012, Ratinov and Roth 2012) and deep neural network solutions (Lee et al. 2017, Zhang et al. 2018), with the best results presently achieved for English.

The chapter concludes with a short summary of evaluation methods and metrics currently used: MUC (Vilain et al. 1995), B^3 (Bagga and Baldwin 1998), CEAF-E (Luo 2005) and CoNLL (Pradhan et al. 2012) together with the two types of clustering algorithms: *mention-pair* (Aone and Bennett 1995) and *entity-based* (Luo et al. 2004).

3 Model of reference and its corpus representation

Chapter 3 presents the model of reference composed of a precise definition of mention types and scope, followed by a broad taxonomy of referential relations. Mentions are defined as generalized nominal groups (nouns with their syntactic constituents, personal pronouns, demonstratives introducing non-relative clauses, elided subjects, gerunds), possessive pronouns and undefined, negative or universal pronouns (often forming pseudoconferential clusters). The schema involves marking nested phrases and discontinuities.

The taxonomy of referential relations (see Fig. 1) is supplemented with the concept of facets representing subjectivity, uncertainty or impartiality of the parties involved in communication. Additionally, the typology presents non-referential auxiliary relations used to support the process of decoding reference.

Chapter 4 presents the process of construction of the corpus of reference relations. After establishing the annotation strategy and sampling texts from the National

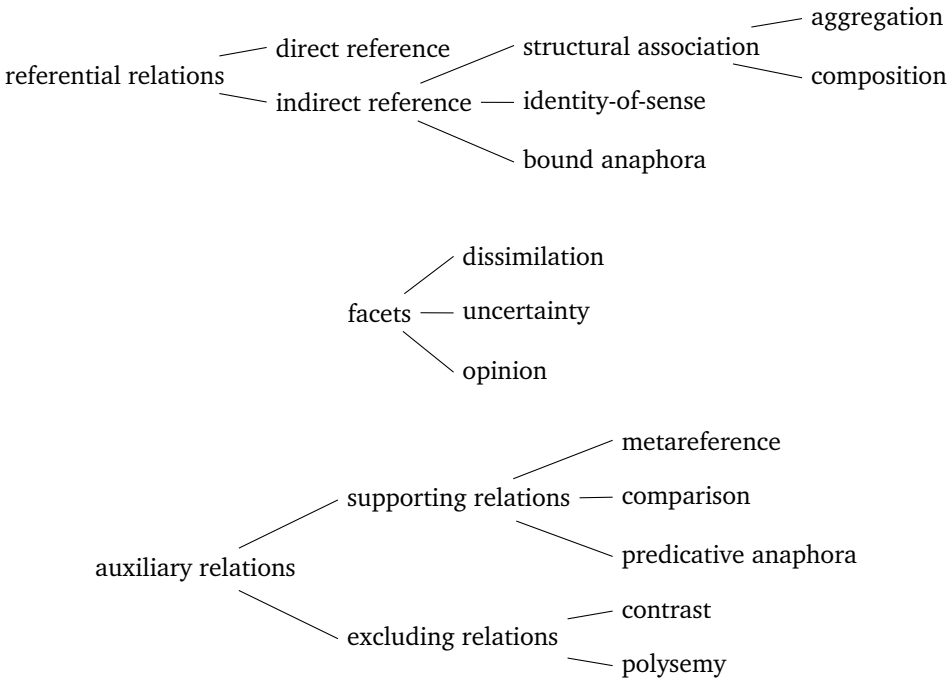

Figure 1. Referential relations, facets and auxiliary relations

Corpus of Polish (Przepiórkowski et al. 2012) the data set was manually annotated, following the designed typology and made available to download, browse and search. As a result, the Polish Coreference Corpus with 530K tokens was created — currently one of the largest coreference corpora in the world. Its basic statistics are presented in Table 1.

4 Implementation and evaluation

Chapter 5 describes the development of the mention detection and coreference resolution tools. Two mention detectors were implemented, both using input from various natural language processing components for Polish: a morphological analyser, named entity recognizer, and a shallow parser. The rule-based detector collected mention candidates from available sources and removed redundancies. The machine-learning detector used a number of lexical, grammatical and numerical features.

Table 1. Basic statistics of the Polish Coreference Corpus

Tokens	532,166
Mentions	185,802
nominal groups	137,645
named entities	20,847
zero subjects	18,807
personal pronouns	9,185
ellipses	488
Clusters	21,865
Bridging relations	
aggregation	9,213
composition	2,008
bound anaphora	317
identity-of-sense	224
Supporting relations	
predicative relation	2,350
metareference	154
comparison	101
Excluding relations	
contrast	225
polysemy	2
Facets	
no facet	16,455
dissimilation	336
opinion	188
uncertainty	66

Implemented coreference resolvers used even more decoding methodologies. The baseline rule-based system was built over a number of manually created rules (gender-number match, nesting prevention, same surface form agreement, etc.), taking into account compatibility of mention pairs and clusters. The machine-learning system used 147 mention-pair learning features representing the surface, syntactic, semantic, meta-textual and anaphoric information. The sieve-based system ordered the classifiers according to their precision and contained six simple compatibility rules. The neural system used word embeddings as components of feature vectors to train the deep neural network model.

Chapter 6 presents detailed results of the evaluation of implemented components over the test portion of the corpus, following the 10-fold cross-validation method. Mention detectors were evaluated in two settings: by comparing only semantic heads of mentions and by comparing the complete borders of mentions (see Table 2).

Table 2. Evaluation of mention detectors

Approach	**Semantic heads**	**Exact boundaries**
Rule-based	89.12%	69.10%
Machine-learning	**91.23%**	**71.79%**

Coreference resolvers were evaluated using traditional metrics in five variants, taking into account gold mentions only or system mentions in two main configurations with respect to mention borders (semantic heads only or complete borders) and the strategy of processing twinless mentions: INTERSECT (considering only mentions present in both gold and system sets) versus TRANSFORM (following the procedure by Pradhan et al. (2011), used in CoNLL-2011 shared task). Table 3 presents the values of CoNLL F_1 measure for all implemented systems.

Table 3. Evaluation of implemented coreference resolution systems

Approach	**Gold mentions**	**Semantic heads**	**Exact boundaries**	**Semantic heads**	**Exact boundaries**
		INTERSECT		**TRANSFORM**	
Rule-based	74.10%	77.05%	78.86%	72.19%	68.88%
Machine-learning	80.50%	80.00%	82.71%	74.82%	**73.96%**
Sieve	80.70%	80.85%	82.49%	75.55%	73.21%
Neural	80.59%	80.89%	**82.73%**	**76.03%**	73.39%
Hybrid	**81.09%**	**81.04%**	82.54%	75.74%	73.17%

5 Reference in discourse

Chapter 7 presents several experiments investigating reference in a broader context of discourse relations. First was the comparison of the description of Prague Discourse Treebank-compatible reference relations in Czech, Russian and Polish with the newly implemented Parallel Annotated Wall Street Journal corpus (Nedoluzhko et al. 2018). The results show variation of referential properties in different languages, both in frequency of the use of referential groups and in types of reference.

Similarly, annotation of discourse relations, including coreference, with the Penn Discourse Treebank methodology for English, German, Polish, Portuguese, Russian and Turkish was carried out, showing differences in realization of discourse relations in different languages (Zeyrek et al. 2019).

Another task focused on annotation of the Polish Coreference Corpus with event-linking time relations, communication events and relations between questions and responses to analyze explicitness and implicitness of representations of events

in the text. The results also show variation in coverage of the text with different metatextual relations which may help investigate how reference influences textual coherence and cohesion.

6 Conclusions

The concluding chapter summarizes the most important findings from the work. The presented study constitutes the first attempt of computer-based large-scale analysis of the nominal referential relations in Polish. The construction of a large corpus and decoding tools made it possible to achieve for Polish the results comparable with global developments. This was made possible by applying a series of improvements on many linguistic and technical levels, starting with the clarification of the notion of reference, anaphora and coreference, through reconstruction of the formal grammar of Polish, integration of external resources and development of new detection algorithms.

Our research, with standardization of the Polish terminology in the field of coreference, the proposed categorization of text-based reference markers and typology of referential relations contributed to the description of the problem in Polish computational linguistics. Creation of one of the world's largest representative corpora of referential relations manually annotated with coreference and bridging relations based on the texts of the reference corpus required proper selection of texts, preparation of annotation guidelines, adaptation of tools and development of the annotation methodology. Several annotation strategies were tested, demonstrating the usefulness of serial adjudication in complex semantic tasks.

The environment for annotation and presentation of the coreference data was prepared with several corpus representation formats, visualisation of referential relations and a search engine linking reference with other layers of linguistic analysis. Rule-based and machine-learning mention detectors were implemented, as well as several solutions for decoding coreference — rule-based, statistical, neural, sieve and hybrid systems, tested in various configurations and supported with external resources, such as the database of periphrastic expressions, valency dictionary or customized formal grammar of Polish. A prototype configuration for detecting associative relations was carried out to decode aggregation, composition and predictive relations. The evaluation of the resulting systems has been conducted in accordance with the commonly used metrics and methodology. In addition, qualitative analysis of created decoders was performed to reduce errors.

In cooperation with foreign partners, we have started cross-lingual research on reference relations linking the work on Polish with other languages. These activities constitute the first step towards the universal multilingual description of coreference.

The results of the presented work are also practical: they have been used, among others, in the automatic summarization system to improve text fluency by means of mention substitution.