Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE

Jan Václ

Sledování aktivovanosti objektů v textech

Ústav formální a aplikované lingvistiky

Vedoucí diplomové práce:	Mgr. Barbora Vidová Hladká, Ph.D.
Studijní program:	Informatika
Studijní obor:	Matematická lingvistika

Praha 2014

I would like to thank Mgr. Barbora Vidová Hladká for her great help and assistance with this thesis and my parents for their admirable support and patience.

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In .... date .... Jan Václ

Název práce: Sledování aktivovanosti objektů v textech

Autor: Jan Václ

Vedoucí bakalářské práce: Mgr. Barbora Vidová Hladká, Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt: V kontextu analýzy diskurzu stupeň aktivovanosti (salience) modeluje aktuální míru zapojenosti odkazovaných objektů a její vývoj v průběhu textu. Algoritmus pro sledování aktivovanost a vizualizaci jejího průběhu již byl navržen a otestován na malém vzorku dat. Tato práce reprodukuje výsledky algoritmu ve větším měřítku pomocí dat z Pražského diskurzního korpusu 1.0. Výsledky jsou pak zpracovány do přístupného tvaru a je provedena jejich analýza jak pomocí vizuálního výstupu, tak i výstupů kvantitativních. Přitom jsou zohledněny dva základní stavební kameny aktivovanosti; koreferenční vztahy a informační struktura věty. V závěru jsou provedeny experimenty zkoumající možné využití informace o aktivovanosti v některé z úloh strojového učení při zpracování přirozeného jazyka na příkladech shlukování dokumentů a tematických modelů.

Klíčová slova: aktivovanost, salience, koreference, TFA, strojové učení

Title: Tracing Salience in Documents

Author: Jan Václ

Supervisor: Mgr. Barbora Vidová Hladká, Ph.D., Institute of Formal and Applied Linguistics

Abstract: The notion of salience in the discourse analysis models how the activation of referred objects evolves in the flow of text. The salience algorithm was defined and tested briefly in an earlier research, we present a reproduction of its results in a larger scale using data from the Prague Discourse Treebank 1.0. The results are then collected into an accessible shape and analyzed both in their visual and quantitative form in the context of the two main resources of the salience – coreference relations and topic-focus articulation. Finally, attempts are made with using the salience information in the machine learning NLP tasks of document clustering and topic modeling.

Keywords: salience, coreference, TFA, machine learning

Contents

Introduction **1**

0.1 Motivation . . . 1

0.2 Goals and Contents . . . 1

0.3 Content Overview . . . 1

1 Research **3**

1.1 Related Work . . . 3

1.2 Salience . . . 4

2 Theory **5**

2.1 Coreference . . . 5

2.1.1 Grammatical and Textual Coreference . . . 6

2.1.2 Bridging Anaphora . . . 6

2.2 Topic-Focus Articulation . . . 7

2.3 Salience . . . 9

2.3.1 Salience algorithm . . . 9

2.4 Latent Dirichlet Allocation . . . 10

2.4.1 Evaluation . . . 12

3 Data and Tools **14**

3.1 Data Sources . . . 14

3.1.1 PDT 2.0 and 2.5 . . . 14

3.1.2 PDiT 1.0 . . . 14

3.1.3 PDT 3.0 . . . 15

3.2 Training and Test Datasets . . . 15

3.3 Tools . . . 17

3.3.1 Tools for PML . . . 17

3.3.2 Topic Modeling Tools . . . 17

3.3.3 Miscellaneous . . . 18

4 Salience Analysis and Interpretation 19

4.1 Sentences, Coreference and TFA Statistics 19

4.1.1 General and Sentence Statistics 19

4.1.2 Coreference 19

4.1.3 TFA 23

4.2 Salience Graphs and Interpretation 24

4.2.1 Salience Graphs 24

4.2.2 Salience Graph Generation Procedure 26

4.2.3 Vertical Cut 28

4.2.4 Horizontal Cut and Leap Height 28

5 Topic Modeling Experiments 32

5.1 Preliminary Experiment – Clustering 32

5.1.1 Sorting the Nouns by Counts 32

5.1.2 Sorting the Chains by Average Salience 32

5.1.3 Clustering Visualization 33

5.2 Latent Dirichlet Allocation and Salience 35

5.2.1 LDA without Salience 35

5.2.2 LDA with Salience 36

5.3 Performance measures 37

5.4 Results 38

5.4.1 Discussion 38

Conclusion 40

Bibliography 42

List of Figures 43

List of Tables 44

Appendix – CD-ROM Contents 45

Introduction

0.1 Motivation

Discourse can be viewed as a sequence of utterances referring to a set of real-world objects. By modeling the dynamic appearance of these references throughout the text, one can acquire a new knowledge about the structure of the text, the importance of these objects in relation to the text, or even the nature of the text. This knowledge can be subsequently used for further investigation in the field of discourse analysis (for example for comparison of the discourse dynamics between two different languages), as well as enhancing the efficiency of an NLP application working with the whole bodies of texts (text segmentation, topic modeling, information retrieval).

0.2 Goals and Contents

There are two main goals of this work. The first one is to investigate more deeply the notion of *salience* as it is defined in (Hajičová et al., 2006). This includes reproducing the experiment described there on a larger amount of data (using the newly available Prague Discourse Treebank), generating the results in a human-examinable form, and analyzing them especially from the quantitative point of view.

The second goal is to examine the salience and its usefulness as an additional feature for an NLP application. Since the type of information the salience brings is closely associated with the topic or subject of the text, the topic modeling was chosen as the exemplar NLP application.

0.3 Content Overview

The chapters in this thesis are ordered roughly from the theory to experiments and results, corresponding to the order in which the underlying work was undertaken. In Chapter 1, the reader is introduced to the research context of our topic, being acknowledged with the related works in the fields we reach into. References to the historcal research for the *salience* notion are presented together with works concerning its main building blocks, *coreference* and *topic-focus articulation*; as well as topic modeling approaches and applications exploiting similar linguistic information.

Chapter 2 presents a necessary overview of both the linguistic theories behind this work and the statistical foundation of the topic-modeling method used and its possible evaluation. Although this overview is not intended to be exhaustive and large details, it should provide the reader with the knowledge needed to understand this work and the presented results, along with directions to further reading if he will be more interested in any of the subjects.

Data and tools used during the experiments and other parts of the work are enlisted and described in Chapter 3, each with a brief information of how and when they contributed.

Perhaps the main part of this work is described in Chapter 4, where the results of the automatic salience analysis on the larger amount of data are presented, both quantified and visualized. These results are preceded by more general statistics of the data in question, providing the necessary context for their better interpretation.

Chapter 5 describes two series of experiments performed to assess the possible contribution of *salience* as a feature to a machine learning application. The first one is an attempt of a simple document clustering based on the importance of words, the second one is a comparison of topic modeling evaluation featuring a popular statistic algorithm.

Contents of the enclosed CD-ROM are described in the Appendix, along with a brief information of each piece of the data and directions how to approach them. Both the scripts and the results contained on the CD-ROM are an integral part of this thesis and represents an important amount of work done within its scope.

1. Research

1.1 Related Work

Several approaches to the analysis of a discourse structure with its dynamic development in relation to a sentence structure can be found among the linguistic theories. Most of them are based on distinguishing two main semantic types of information in the sentence: *given* vs. *new* (although their terminology varies, not necessarily according to possible differences in the definitions).

Hajičová (2013) mentions another interesting approach to relating the sentence structure with a dynamicity of the discourse structure, given by (Prince, 1981); a three-level hierarchy of *givenness* of an information (contrasting the given-new) between speaker and hearer is presented there. Each level refers to a different understanding of givenness in the works of previous researchers:

1. givenness as a predictability/recoverability, as defined by (Kuno, 1972) and (Halliday, 1967) (althour their definitions slightly differ),
2. givenness in the sense of saliency, relating to the assumption of the hearerś consciousness, referring to (Chafe, 1976),
3. givenness in the relation to a state of a "shared knowledge" according to (Susan E. Haviland, 1974), focusing on what the hearer "already knows and accepts to be true" vs. what the hearer "does not yet know".

Prince then continues with defining a more fine-grained familiarity scale on discourse entities.

Another well-known approach of modeling discourse dynamics in terms of sentence structure is the *centering* theory introduced in (Joshi and Weinstein, 1981) and further refined in (Grosz et al., 1995), based on the local attentional states of speaker and hearer. It operates with a forward and backward looking centers of sentences and defines four types of sentence transitions by the relations of their centers. One of the characteristic features of this theory is ranking of the centers according to a language-specific parametrization.

An entity-grid model is presented in (Barzilay and Lapata, 2008), where each entity appearing in a text (based on a coreference relations) is assigned a column in a grid, each sentence corresponds to a row in this grid. The cells are then filled with syntactic roles of the entities in the corresponding sentence, recording also the transitions between those sentences. It should be noted that this approach, among all the mentioned so far, is the most computationally oriented. Distributional information about the entities are extracted naturally from the entity-grid as well, forming the parameter of *salience* as a discourse prominence. However, our understanding of this notion is slightly different, arguing with Hajičová (2013) that it should be understood in a more complex way and that neither frequency nor the length of the referential chain is a sufficient measure of salience.

Even more application-oriented approach is presented in (Sauper et al., 2010), building a statistical-based model of content structure for using it in a text analysis. This model combines hidden Markov models and conditional random fields, employing the expectation-maximization technique for finding their parameters.

1.2 Salience

Our approach directly follows the notion of *salience* first mentioned and described in (Hajičová and Vrbová, 1982), revisited in (Hajičová, 2003) and further refined and tested in (Hajičová et al., 2006). This notion relates the dynamicity of the discourse with the information structure of its individual sentences, working with activation of the elements of knowledge shared between the speaker and the hearer.

2. Theory

The theory of *salience* will be introduced in Section 2.3, but first, one has to understand two main resources standing behind this notion: coreference and topic-focus articulation.

2.1 Coreference

Coreference is a concept describing a relation of two or more expressions in a text referring to the same real-world object. These expressions are called *referents*.

The key approach to coreference in this work is that the groups of coreferents join together to form a *coreference chain*. When speaking about two neighbouring members of the coreference chain and their relation, the first one is often called the *antecedent* and the second one the *anaphor*, with respect to the order of their occurence in the text. These terms describes the most typical form of the coreference called *anaphora*, when the first expression is the more specific one and the second one relates to the first one – when visualizing the coreference relations, this is often denoted by an arrow directed from the second one to the first one. The reverse case, called *cataphora* is also possible; however, the terminology differs here, the “target” of the relation is usually denoted as *cataphor* and it is now preceded by the “source”, which is often called the *postcedent*.

The distinction between anaphora and cataphora is illustrated by the simple examples (1) in Czech (constructed based on our language experience). The English translations are as literal as possible to retain the structure of the original sentence. The coreference pairs in both cases are highlighted and subscripted.

(1) a. **Krabice**$_{\mathbf{a}}$ byla tak těžká, že **ji**$_{\mathbf{a}}$ Petr raději nechal za dveřmi.
The-box$_{\mathbf{a}}$ was so heavy that **it-OBJ**$_{\mathbf{a}}$ Peter-SUBJ rather left behind the-door.

b. Ačkoliv **ho**$_{\mathbf{c}}$ nikdo nezval, **Martin**$_{\mathbf{c}}$ se-objevil na každém večírku.
Although **him**$_{\mathbf{c}}$ no-one invited, **Martin**$_{\mathbf{c}}$ showed-up on every party.

2.1.1 Grammatical and Textual Coreference

According to the approach to coreference captured in the Czech dependency treebanks and described e.g. in (Kučová and Hajičová, 2004) (with its extension in (Nedoluzhko, 2011)), we distinguish two types of coreference relations in this work, *grammatical* and *textual.* The grammatical coreference in this approach is such a kind of coreference in which it is possible to pinpoint the coreferred expression on the basis of grammatical rules; it may involve a verb of control, reflexive pronouns, verbal complements, reciprocity and relative pronouns. On the other hand, the textual coreference is not realised by grammatical means alone, but also via context. The former type of coreference usually occurs with both the involved coreferents within one sentence, while the latter often cross the sentence boundaries.

2.1.2 Bridging Anaphora

The term *bridging anaphora*, also sometimes denoted as *associative anaphora*, is used in this work in correspondence to its annotation in the Prague Discourse Treebank[1], described in detail in (Nedoluzhko, 2011). The term describes an anaphoric relation where the anaphor is not directly coreferential to the antecedent, but an indirect connection is implied. This connection can be identified by the reader often using a real-world knowledge and a cognitive process, sometimes also based on the context. As it is shown in (2) (taken from (Nedoluzhko, 2011)), some knowledge of semantic structures of the mentioned object has to be employed to recognize the relationship between *classroom* and *children.*

(2) **Učitel** vešel do **třídy**. Děti (se) okamžitě přestaly bavit.
(**Teacher** entered (to) **the-classroom**. **Children** instantly stopped talking.)

Within the notion of *bridging anaphora*, more specific subtypes of relations are distinguished, corresponding to the semantic relation of the two referred objects. Based on a rigorous research and analysis of the impact of the inter-annotator agreement, (Nedoluzhko, 2011) settles for the following six subtypes for the Prague Discourse Treebank annotation task:

1. part-whole relation (asymmetric, with both possible directions)
   – e.g. "room"-"ceiling", "finger"-"hand"

[1] For details on the treebank, see 3.1

2. set-subset relation (asymmetric, with both possible directions) – e.g. “drinks”-“beer”, “drinks”-“soda”

3. functional relation (asymmetric, with both possible directions) – e.g. “coach”-“team”, “company”-“director”

4. semantic or pragmatic contrast (symmetric), depends heavily on the context – e.g. “**Last year** we went abroad on holiday, but **this summer** we are staying at home.”

5. non-coreferential anaphoric relation (symmetric) – e.g. “**Love**? What does **the word** even mean?”

6. other – intended for collecting specific types of relations, possibly detachable into their own category in the future: family membership, place-inhabitant, author-piece, possession-owner etc.

Although some of the bridging relations are inherently asymmetric, the members of the anaphoric chain are considered to be equivalent. Thus, we can actually speak of *chains*, with each member referring to the directly previous one.

2.2 Topic-Focus Articulation

Information structure of a sentence is an important aspect of the sentence meaning, especially in the perspective of a discourse analysis. Our understanding of the sentence information structure is directly based on the Functional Generative Description framework (FGD), i.e. the approach of the Prague School of Linguistics. An insightful survey of this approach can be found in (Hajičová, 1993), for more detailed treatment see eg. (Sgall et al., 1986).

The key notion in this approach is the *topic-focus articulation*[2] (or TFA), a partioning of the sentence into two segments each with different communicational function. In the *topic* part of the sentence, the speaker mentions “what he is talking about”, while the *focus* part contains new information about the topic, i.e. “what he wants to say about it”. The dichotomy links the semantic structure of a sentence with the structure of discourse in its context, and is usually found to be also anchored in the syntactic structure of the sentence. Natural languages use various surface means to convey this distinction: word order plays the main role in inflectional languages, specific morphemes are present in several languages of

[2] This dichotomy is sometimes described also as *theme/rheme*, *topic/comment* or *presupposition/focus* by more traditional theories and also by similar contemporary approaches. However, the main distinguishing principles rarely differs.

Eastern Asia, e.g. in Japanese, and intonation seems to be important everywhere, espedal1y in the analytic languages of Western Europe; German combines in various respects the properties of the latter with these of inflectional languages (Hajičová, 1993).

An example sentence in Czech (from (Hajičová et al., 2005)) is shown in (3) to illustrate the topic-focus segmentation.

(3) V noci ze soboty na neděli skončil ve vojenském prostoru Ralsko sjezd majorů.
(At night from Saturday to Sunday ended in military area Ralsko meeting(Nom.) of-majors.)
Topic: v noci ze soboty na neděli (at night from Saturday to Sunday)
Focus: skončil ve vojenském prostoru Ralsko sjezd majorů (ended in military area Ralsko meeting(Nom.) of-majors)

As stated in (Hajičová et al., 2005) and following the FGD approach, the semantic basis of the articulation of the sentence in to Topic and Focus is the relation of contextual boundness: a prototypical declarative sentence asserts that its Focus holds (or does not hold) about its Topic. Within both Topic and Focus, an opposition of contextually bound and non-bound nodes is distinguished, which is understood as a grammatically patterned opposition, rather than in the literal sense of the term. Within the contextually bound elements of the sentence, a difference is made between contrastive and non-contrastive bound elements.

Following the theoretical assumptions of FGD, TFA is captured in the tectogrammatical annotation of the Prague Dependency Treebank[3] by the TFA attribute, which may obtain one of the three values:

- *t*: a non-contrastive contextually bound node,
- *c*: a contrastive contextually bound node,
- *f*: a contextually non-bound node.

Returning to the relation of the two different views, the semantic view represented by the contextual boundness and non-boundness serves as a basis for inferring the syntactic, surface-form Topic/Focus dichotomy and possible segmentation of a sentence. In this direction, a heuristic procedure was proposed by (Sgall et al., 1986) to identify the sentence bipartition of Topic/Focus based on the distinction of contextually bound and non-bound items.

[3] For details on the treebank, see 3.1

2.3 Salience

The flow of a discourse can be viewed as a sequence of sentences, each with its own information structure and most of them referring to some real-world objects. In different parts of the discourse, some of these objects are referred to more often than the others and vice versa. The notion of *salience* suggests that at every point of the discourse, i.e. in every sentence, each of these objects can be assigned a certain level of *activation*, or *salience*.

One can assume that all the objects referred in a discourse are taken from some *stock of knowledge* shared between the speaker and the hearer (or, in case of a written text, the author and the reader). Then we can regard this set of objects rather as a stack, bearing the most activated items on the top. When an object is mentioned in a sentence, it is moved to the top of the stack (or closely to it, depending on the usage of the referring expression in the sentence). Then, if not referred in the following sentences, it slowly descends, pushed down by the objects which are mentioned in these sentences. Given this model, the quantity of *salience* of an object determines how high this object is located on the stack.

Assumptions have been made (Hajičová, 2003) that if the salience values of the referenced objects in a discourse could be determined, one would be able to induce various characteristics of the discourse. One of them is observing a segmentation of the discourse according to groups of momentarily salient objects along with the identification of their topic(s). Another one could be prediction of a grammatical form of the referring expressions (or, more generally, their strength), eg. pronominal vs. noun referent. Some of these assumptions will be addressed and analyzed in this work.

2.3.1 Salience algorithm

A deterministic procedure to determine the salience values of the coreference chain in the flow of a discourse on a sentence-by-sentence basis in (Hajičová et al., 2006). Its evaluation was presented on one sample document only, because not much data with the necessary annotation were conveniently available at that time. However, the results of the algorithm were also visualized to provide more human-readable feedback.

Let us recall the salience algorithm, as defined in (Hajičová et al., 2006) – consider the following situation: An object x represented by the referent r has the salience degree $dg_x^n(r)$ after the n-th sentence of a document is uttered. Then, the salience value of the object x is defined after its first mentioning by the linear sentence-by-sentence processing as follows:

After each sentence, the salience degree of the object x is modified:

1. $dg_x^n(r) = -1$ if r carries TFA value t or c in the n-th sentence,
2. $dg_x^n(r) = 0$ if r carries TFA value f in the n-th sentence,
3. $dg_x^n(r) = dg_x^{n-1}(r) - 2$ if r is not included in the n-th sentence and has been mentioned in the Focus of the last (not necessary immediately) preceding sentence ($(n-1)$-th through 1st sentence),
4. $dg_x^n(r) = dg_x^{n-1}(r) - 1$ if r is not included in the n-th sentence and has been mentioned in the Topic of the last (not necessary immediately) preceding sentence ($(n-1)$-th through 1st sentence).

Note that this formulation of the salience algorithm does not define the salience value of x before it is first mentioned in the document.

The salience algorithm distinguishes between the Topic/Focus dichotomy and the TFA attribute values ($c/t/f$), according to the theoretical background summarized in 2.2. However, in the scope of this work, we will make a simplification at this point and use the notion Focus synonymously to the TFA value f and likewise Topic synonymously to the TFA values c or t. The reasons are rather of technical nature; although a heuristic algorithm proposed by (Sgall et al., 1986) has been stated and tested in (Hajičová et al., 2005) for “converting” the $c/t/f$ values to Topic/Focus, its results were not fully deterministical. Furthermore, this algorithmic procedure could not be reproduced within the scope of this work.

2.4 Latent Dirichlet Allocation

Currently one of the best known and very broadly used methods for topic modeling tasks is Latent Dirichlet Allocation (or simply LDA). It was introduced in (Blei et al., 2003) as a generative probabilistic model for collections of discrete data, such as text corpora. The model has three layers: the items of such collection are modeled as a finite mixture over an underlying set of topic probabilities, each topic modeled as an infinite mixture over an underlying set of topic probabilities; the topic probabilities provide an explicit representation of a document.

The output of the model is a given number of topics, each of which, as mentioned earlier, is a defined by a list of probabilities over a set of words. Thus,

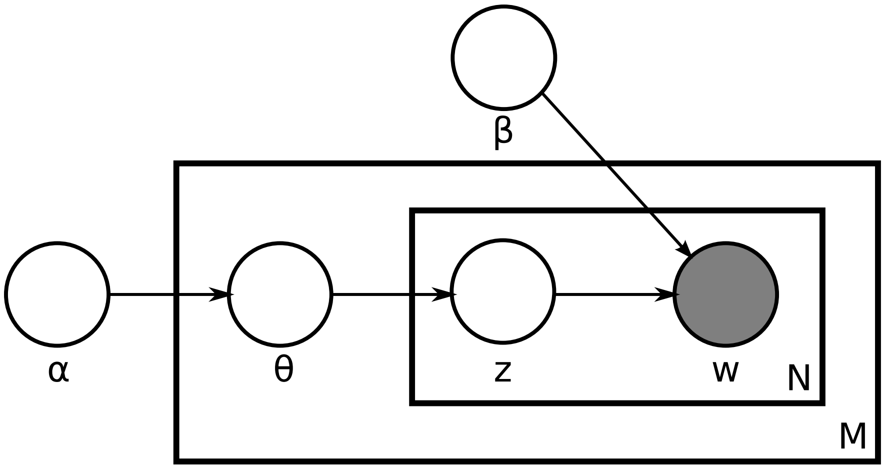

Figure 2.1: Graphical model representation of the unsmoothed LDA model in a plate notation (source: (Blei et al., 2003)).

no single representant or a name for a topic is inferred, a “meaning” of a topic is present only inherently, by a collective contribution of words with higher probabilities. Also, no word is exclusive for any topic, its “assignment” is always proportionally divided between more of them.

The plate notation of the LDA model in Figure 2.1 summarizes its three-layer architecture with the probabilistic distributions and their corresponding parameters. The boxes are “plates” representing replicates. The outer plate represents documents, the inner plate represents the repeated choice of topics and words within a document. M denotes the number of documents, N the number of words in a document. The parameters are then as follows:

- α is the parameter of the Dirichlet prior on the per-document topic distributions,
- β is the parameter of the Dirichlet prior on the per-topic word distribution,
- θ_i is the topic distribution for document i,
- ϕ_k is the word distribution for topic k,
- z_{ij} is the topic for the jth word in document i, and
- w_{ij} is the specific word.

LDA assumes the following generative process for a corpus D consisting of M documents each of length N_i:

1. Choose $\theta_i \sim \mathrm{Dir}(\alpha)$, where $i \in \{1, \ldots, M\}$ and $\mathrm{Dir}(\alpha)$ is the Dirichlet distribution for parameter α
2. Choose $\phi_k \sim \mathrm{Dir}(\beta)$, where $k \in \{1, \ldots, K\}$

3. For each of the word positions i, j, where $j \in \{1, \ldots, N_i\}$, and $i \in \{1, \ldots, M\}$
   (a) Choose a topic $z_{i,j} \sim \text{Multinomial}(\theta_i)$.
   (b) Choose a word $w_{i,j} \sim \text{Multinomial}(\phi_{z_{i,j}})$.

Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is given by:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod p(zn|\theta)p(wn|zn, \beta) \tag{2.1}$$

Inference Learning the various distributions (the set of topics, their associated word probabilities, the topic of each word, and the particular topic mixture of each document) is a problem of Bayesian inference. The original paper used a variational Bayes approximation of the posterior distribution, although more efficient alternative inference techniques exist using a collapsed Gibbs sampling and expectation propagation (Minka and Lafferty, 2002).

2.4.1 Evaluation

The task of evaluating an LDA model is a general task of a generative model evaluation: we want to compute the probability of a held-out document collection given this trained model. In particular, we want to maximize the probability:

$$P(W|\phi, \alpha m) = \prod_d P(w^{(d)}|\phi, \alpha m) \tag{2.2}$$

Since the topic assignments for one document are independent of the topic assignments for all other documents, each held-out document can be evaluated separately.

However, the exact computation of this probability is intractable, thus various estimators are used, such as harmonic mean or empirical likelihood method and others. A rigorous experimental comparisons of these estimators is presented in (Wallach et al., 2009) and more accurate alternatives are proposed; a Chib-style estimator and a left-to-right evaluation algorithm. The latter was introduced in (Wallach, 2008) and is observed to be even more accurate on real-world datasets than the former. It is currently commonly used method for the LDA evaluation and was chosen also for the task in this work.

Left-to-right evaluation For each document w, its latent topic assignments z and its document-specific topic distribution θ, we compute:

$$\begin{aligned} P(w|\phi, \alpha m) &= \prod_n P(w_n | w_{<n}, \phi, \alpha m) \\ &= \prod_n \sum_{z \leq n} |w_{<n}, \phi, \alpha m) \end{aligned} \tag{2.3}$$

Each sum over $z \leq n$ can then be approximated using an approach inspired by sequential Monte Carlo methods, the algorithm pseudo-code being as follows:

Algorithm 1 Left-to-right evaluation algorithm

```
initialize $l := 0$
for each position $n$ in $w$ do
    initialize $p_n := 0$
    for each particle $r = 1$ to $R$ do
        for $n' < n$ do
            sample $z_{n'}^{(r)} \sim P(z_{n'}^{(r)} | w_{n'}, \{z_{<n}^{(r)}\}_{\backslash n'}, \phi, \alpha m)$
        end for
        $p_n := p_n + \sum_t P(w_n, z_n^{(r)} = t | z_{<n}^{(r)}, \phi, \alpha m)$
        sample $z_n^{(r)} \sim P(z_n^{(r)} | w_n, z_{<n}^{(r)}, \phi, \alpha m)$
    end for
    $p_n := p_n / R$
    $l := l + \log p_n$
end for
$\log P(w|\phi, \alpha m) \simeq l$
```

3. Data and Tools

3.1 Data Sources

3.1.1 PDT 2.0 and 2.5

The *Prague Dependency Treebank* (PDT) represents the largest annotated corpus of Czech language.[1] The texts are syntactically analyzed using the dependency approach with the main role of the verb. The annotations go from the morphological layer through to the intermediate syntactic-analytical layer to the tectogrammatical layer (the layer of an underlying syntactic structure). The process of annotation was performed in the same direction, i.e. from the simplest layer to the most complex. This fact corresponds to the amount of data annotated on each level – 2 million words have been annotated on the lowest morphological layer, 1.5 million words on both the morphological and the syntactic layer, and 0.8 million words on all three layers.

The format of the files containing the annotated data of the PDT family (since PDT 2.0) is called the Prague Markup Language (PML) and is based on XML. Each document data consists of four XML files (typically compressed), one file with the tokenized documents only, each of the rest corresponding to one layer of the annotation and referencing the layer directly superior. Thus, e.g. the tectogrammatical layer as the deepest one, does not contain any surface word forms or purely morphological information itself, but they are accessible through the references.

In 2012, an updated version of PDT 2.0 was released, denoted PDT 2.5. From the aspects examined in this work, the changes between these two versions were not significant. However, instead of one of these two versions, the treebank directly related to PDT 2.5 was used in this work, the Prague Discourse Treebank.

3.1.2 PDiT 1.0

The *Prague Discourse Treebank 1.0* (PDiT)[2] is an extension upon the PDT 2.0. It represents a new manually annotated layer of language description, above the existing layers of the PDT (morphology, surface syntax and underlying syntax) and it portrays linguistic phenomena from the perspective of discourse structure and coherence. The discourse layer of the treebank contains two subprojects:

[1] http://ufal.mff.cuni.cz/pdt2.0

[2] http://ufal.mff.cuni.cz/discourse

- lexically-grounded approach of identification of discourse connectives, discourse units linked by them and semantic relations between these units;
- annotations of extended textual coreference and bridging relations.

With its 49,431 manually annotated sentences from Czech newspapers, the project serves as a large-scale resource for linguistic research in the area of discourse analysis as well as for computational experiments concerning automatic text analysis, information extraction, text summarization and other branches of NLP research.

Figure 3.1 taken from the Prague Discourse Treebank annotation manual visualizes the tectogrammatical tree structure of one sentence, along with an arrow visualization of the coreference relations. The notation also distinguishes the grammatical and textual reference and includes a bridging anaphora relation. Each tectogrammatical node (or simply *t-node*) has its attributes visualized, such as its tectogrammatical lemma ("potřebovat"), functor ("ACT", "PAT", "PRED",...) or a specific sub-type of its reference relation ("SPEC", "WHOLE_PART"). Also note that there are some t-nodes added without any counterpart in the surface representation – such as the root node of the sentence. Another examples would be technical nodes generated e.g. in places of naturally elided expressions, such as zero pronouns. On the other hand, some surface tokens are not represented in the tectogrammatical structure, such as prepositions or auxiliary verbs, their function in the sentence is captured by the attributes of the existing t-nodes.

The Prague Discourse Treebank is the only source of linguistically annotated data used for the purposes of this work.

3.1.3 PDT 3.0

Shortly before finishing this thesis, a new member of the PDT series was published: the Prague Dependency Treebank 3.0.[3] Compared to PDiT, it has been enriched by new types information such as genre annotation, extension of the textual coreference with 1st and 2nd person pronominals; as well as revised in various aspects. This version of PDT was not yet used in this work, however, it is encouraged by any possible following research to use it.

3.2 Training and Test Datasets

PDiT has already prepared 3 groups of datasets according to the data partitioning typical for the NLP tasks: the training data, the development test data and

[3] http://ufal.mff.cuni.cz/pdt3.0

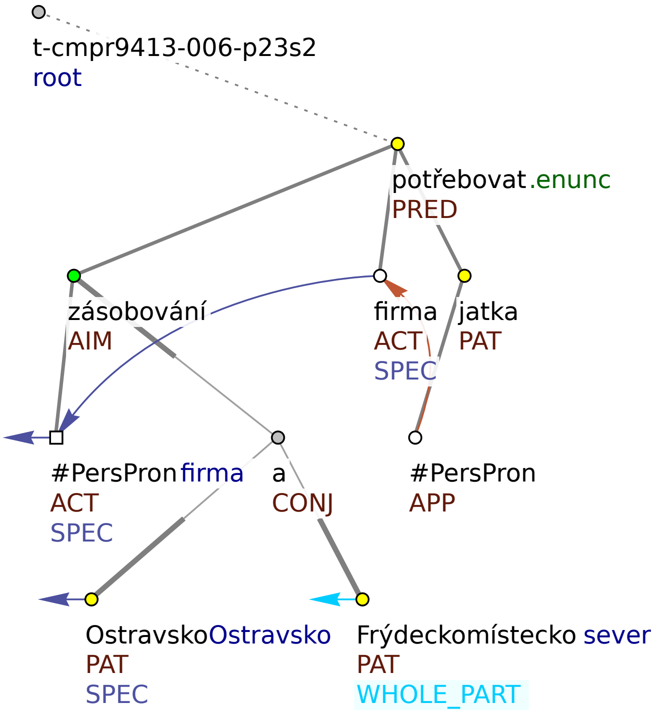

Figure 3.1: Example of coreference annotation for the following sentence: *Pro zásobování Ostravska a Frýdeckomístecka potřebuje firma svá jatka. (The company needs its slaughterhouse in order to supply the Ostrava and Frydek-Mistek regions.)* The dark red arrow is used for a grammar coreference relation, dark-blue arrows for textual coreference; light-blue arrow for bridging reference (to an expression in another sentence).

the evaluation test data. The training datasets cover approximately 80%, development 10% and evaluation 10% of the whole set of data (these proportions hold for all the three layers of annotation).

In this work, we exploit the prepared partitioning of PDiT, but we do not use the evaluation data at all. Furthermore, for preliminary experiments and some of the more time-consuming tasks, we use only one eighth of the whole training data, the part denoted train-1. Throughout this work, we will often refer to this smaller subset as *train-1*, in contrast to the whole training set, denoted as *train-all*. The development-test data, used for the evaluation of the experiments, will be referred to as *dtest*.

For a more detailed quantitative analysis of the datasets from the perspective of the features investigated in this work, see Section 4.1.1.

3.3 Tools

3.3.1 Tools for PML

For the batch-processed salience analysis, more convenient data browsing and other manipulation, several tools were used:

- `btred`[4] – Perl-based interface for macro scripting specialized on processing the PML data. Created as a tool for PDT 2.0 (thus applicable also on PDiT), and used in this work especially for various data format conversions.
- `Tree Editor TrEd`[5] – a viewer and editor of the PDT annotation files, part of the PDT 2.0 distribution. Additional plugins were installed for handling the extra attributes, e.g. color of the coreference strings.
- `XSH2`[6] – XML editing shell, used for the extraction of lemmata from the PDT XML format.

3.3.2 Topic Modeling Tools

- `MALLET`[7] (McCallum, 2002) – MAchine Learning for LanguagE Toolkit, a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other

[4] `http://ufal.mff.cuni.cz/~pajas/tred/btred.html`

[5] `http://ufal.mff.cuni.cz/~pajas/tred/index.html`

[6] `http://xsh.sourceforge.net/`

[7] `http://mallet.cs.umass.edu/`

machine learning applications to text. In this work, it was used for the topic modeling and evaluation.

- `gensim`[8] (Řehůřek and Sojka, 2010) – a free Python library for scalable statistical semantic analysis and topic modeling. Its implementation of LDA was used in preliminary experiments with topic modeling.

3.3.3 Miscellaneous

- `R`[9] – the R language for statistical computing was used for plotting the salience graphs.

- `Perl`[10] programming language – used for simpler text and data manipulation.

- `Python`[11] programming language – used for various more complicated data manipulation, as well as for plotting some of the bar charts.

- `pygraphviz`[12] – a Python interface to the Graphviz[13] open source graph layout and visualization package. It was used here for the visualization of the preliminary topic modeling experiment.

- `LibreOffice`[14] `Calc` – a spreadsheet program used for manipulating and plotting especially the data of salience leap heights.

- various Unix shell scripts and makefiles – for the smaller tasks, especially for the purposes of batch execution of the tasks, some simple scripts were written for the purposes of this work. These tasks included especially the output evaluation, but also grid-searching for parameters or format conversions and adaptations of the data. All these scripts are also present as a part of this work on the enclosed CD-ROM (see Section 5.4.1).

[8] `http://radimrehurek.com/gensim`
[9] `http://www.r-project.org/`
[10] `http://www.perl.org/`
[11] `http://www.python.org/`
[12] `http://pygraphviz.github.io/`
[13] `http://www.graphviz.org/`
[14] `http://www.libreoffice.org/`

4. Salience Analysis and Interpretation

4.1 Sentences, Coreference and TFA Statistics

Before we proceed to analyze the salience models and its behavior, we should present some statistics about the data and the features which the salience is built upon. Also, the quantitative characteristics of the documents at hand will be useful in the later part, when building and testing the topic models.

4.1.1 General and Sentence Statistics

Table 4.1 presents an overview of general quantitative characteristics for both training sets used further in the experiments.

	train-1	*train-all*
No. of documents	316	2533
Total no. of sentences	4700	38727
Avg. no. of sentence per doc.	14.9	15.3
Total no. of t-nodes	68626	567258
Avg. no. of t-nodes per sentence	14.6	14.6
Avg. no. of t-nodes per doc.	217.2	223.9

Table 4.1: General statistics of the datasets.

More detailed distribution of the sentence counts in documents is shown in Figures 4.1 and 4.2. Note that the most typical sentence count in both cases is 8, which is far below the average value.

4.1.2 Coreference

Perhaps the more important one of the two main pillars which the salience concept is built upon, is the concept of the coreference relation. To understand the salience models, we have to explore first the basic characteristics of the coreference chains themselves in our data.

The counts of the grammatical and textual coreference links in *train-1* and *train-all* are summarized in Table 4.2 and Figure 4.3 along with the counts of bridging anaphora links. Those are not coreference relations in the strict sense,

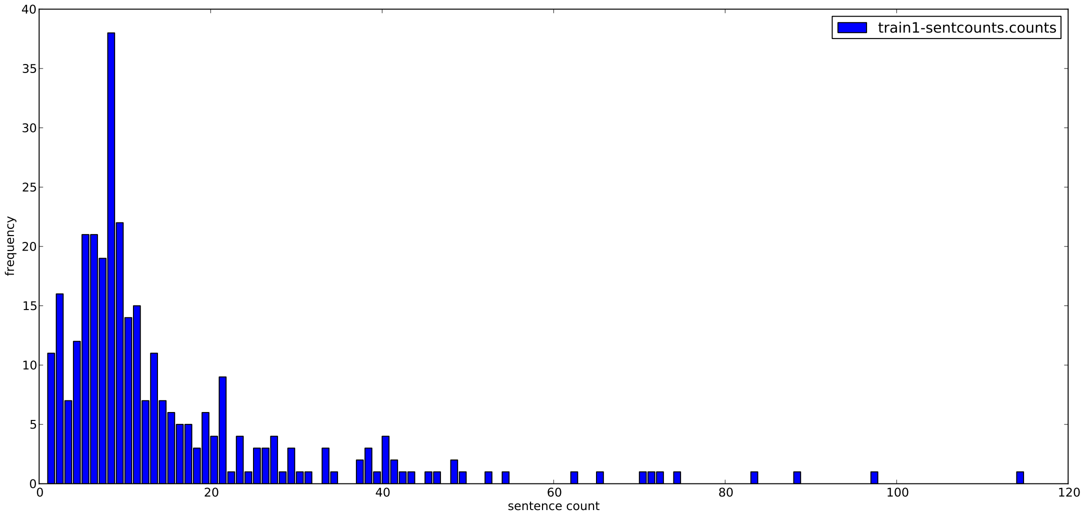

Figure 4.1: Distribution of the per-document sentence counts in *train-1* dataset.

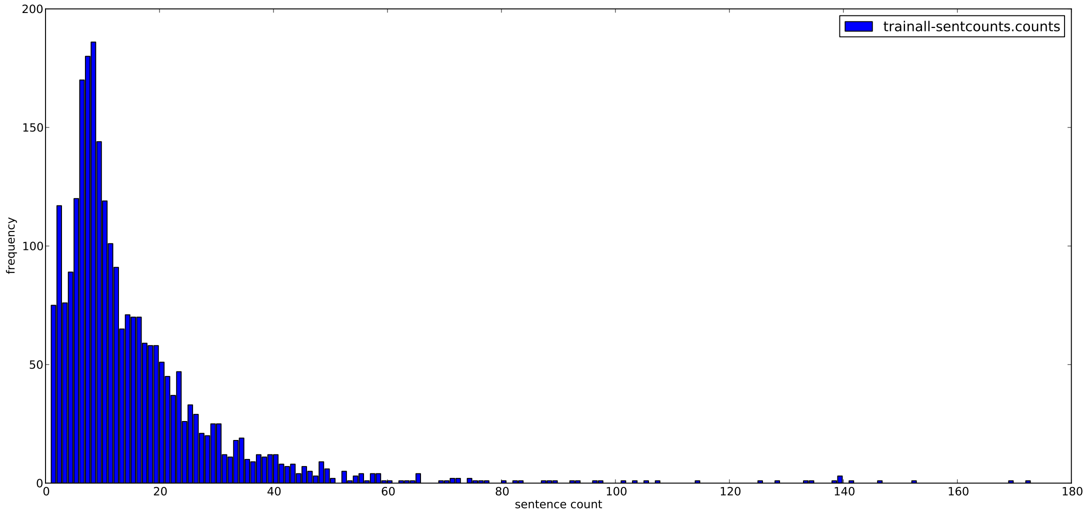

Figure 4.2: Distribution of the per-document sentence counts in *train-all* dataset.

coreference type	*train-1*	*train-all*
grammatical	2226	18156
textual	7514	67535
bridging anaphora	1987	23512

Table 4.2: Coreference type link counts

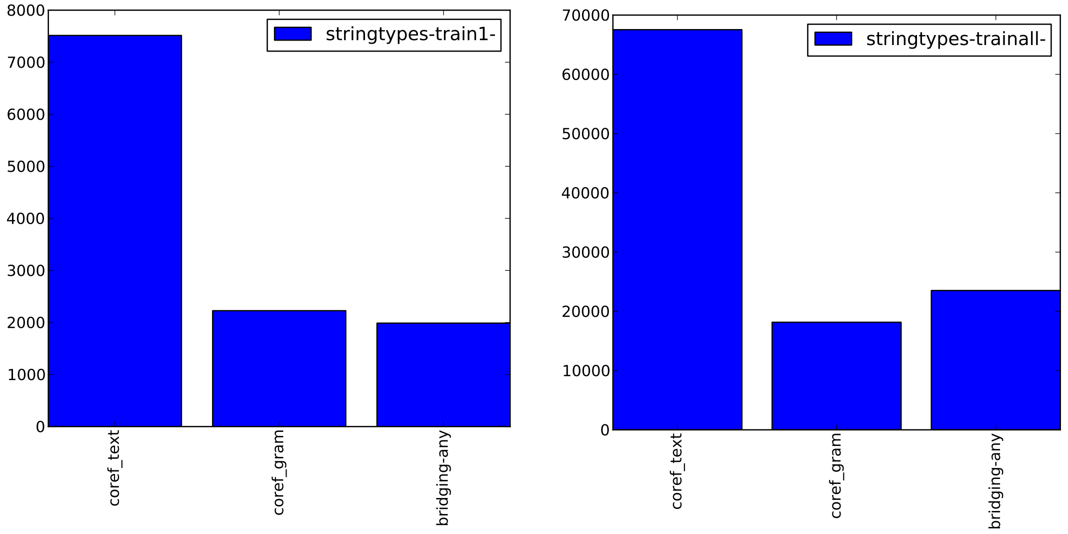

Figure 4.3: Counts of the coreference link types in *train-1* and *train-all* dataset.

but since we have experimented with using them as such (see Section 4.1.2), the numbers are listed there for comparison.

Another important data are the counts of the whole coreference chains, presented in Table 4.3. These frequencies, especially when related to the number of documents, might be crucial in some decisions concerning topic modeling, especially when using bag-of-words models and simulating the word counts with the coreference chains (see further in Chapter 5).

	train-1	*train-all*
No. of documents	316	2533
Total no. of coref. chains	4519	39415
Avg. no. of coref. chains per doc.	14.3	15.8

Table 4.3: Counts of the whole coreference chains in the datasets, related to numbers of documents.

Chain lengths When speaking of the length of a coreference chain, we have adopted the definition of coreference chain length being the number of coreference nodes (i.e. co-referring expressions) in the chain. Thus the most frequently

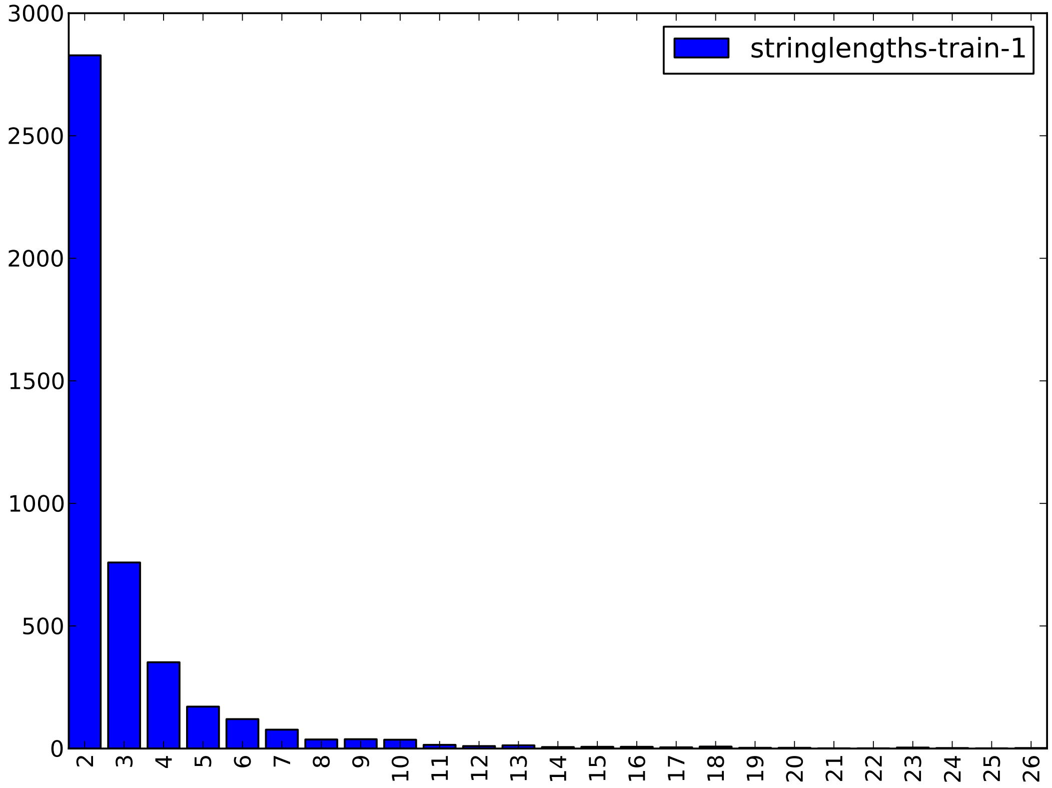

Figure 4.4: Frequency of lengths of coreference chains in *train-1* dataset; cut off at length of 26 nodes.

appearing chain has length of 2, meaning two anaphoric expressions referring to the same item (in the PML representation, this is represented by two tectogrammatical nodes with one coreference relation between them, typically the first one being the antecedent of the second one). According to this definition, we have acquired the length-frequency figures presented in Figure 4.4. The distribution is not suprising; the chain length of 2 coreferents is the most typical case, whereas the frequency of longer chains drops rapidly. However, although the "tail" of the graph was cut off for sake of readability, the longest chain encountered in the data was 89 nodes long (and it was found in a document of 114 sentences). To complete the data, we will add that the average length of a corefence chain in *train-1* is 5.1.

Adding Bridging Anaphora The coreference chains are the main platform for the salience analysis and modeling of a text. If the salience should be used to model the dynamics of some inherent topics of the text, it would be convenient to have at our disposal the coreference chains "as long as possible". In other words, one should make effort to identify as many connecting relations between

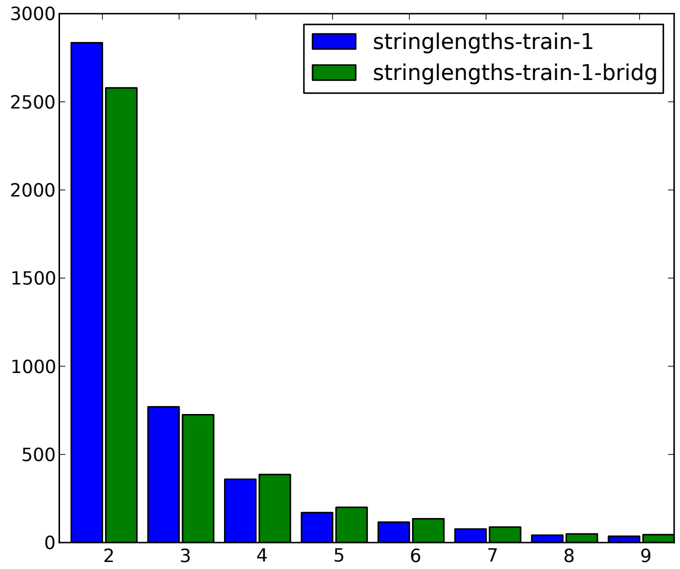

Figure 4.5: Frequency of lengths of coreference chains in *train-1* dataset – the impact of adding bridging anaphora.

associated expressions as possible. In this pursuit, we have experimented also with using the annotation of bridging anaphora as an additional source of coreference relations. The experimental approach was quite straightforward; since the salience algorithm does not distinguish between types of coreference, we can let it treat the bridging relations the exact same way as the "regular" coreference.

However, when commiting to this step, one has to bear in mind that the bridging relations does not have so "strict" characteristics, which can, to a certain degree, also affect the results of the subsequent salience modeling. The measure of this effect can be hardly anticipated – ideally, one would have to perform two sets of all the planned experiments and maintain two sets of results, comparing them and evaluating the differences continuously.

Furthermore, when we examine the actual impact on the length of the coreference chains (see Figure 4.5), the influence is obvious, but not as large as we presumed. Taken into account the above objections, we have finally decided to abandon this path in the scope of this work and perhaps leave it for further investigation.

4.1.3 TFA

The proportion of the TFA markers for the tectogrammatical nodes in *train-1* dataset is visualized in Figure 4.6. In accordance to the PML annotation customs,

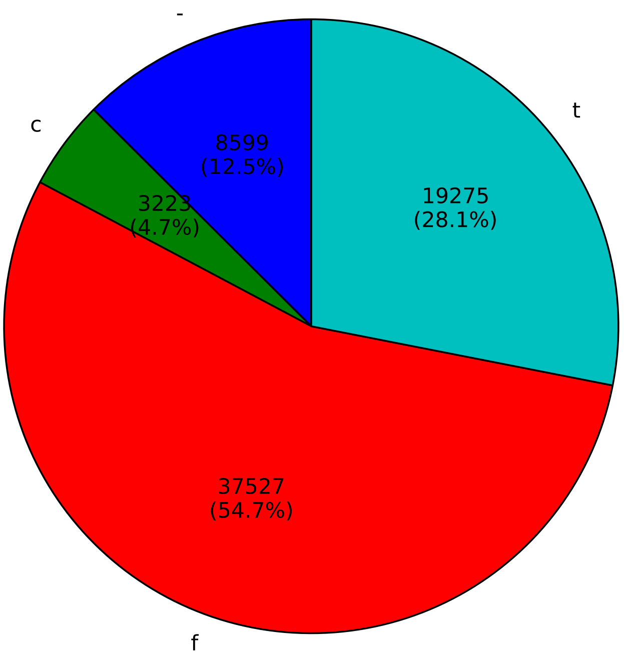

Figure 4.6: Frequency of TFA values in *train-1* dataset.

`t` stands for non-contrastively contextually bound expression (represented by the node), `c` for contrastive contextually bound expression and `f` for contextually non-bound expression. Finally, the `-` part of the chart represents the amount of nodes not marked with TFA values[1].

4.2 Salience Graphs and Interpretation

4.2.1 Salience Graphs

Figure 4.2.1 presents an example of a salience graph for a short document. The graph was generated from the Czech original of the document, the presented English translation tries to preserve partially the original sentence structure. In the chart, each coreference chain is represented by a numbered polyline, the members of the chain are marked by the corresponding color in the text.

[1] These are mostly technical cases, e.g. root of the tectogrammatical tree or of a paratactic construction, or a foreign-language expression, which has often a special treatment in the PML annotation scheme

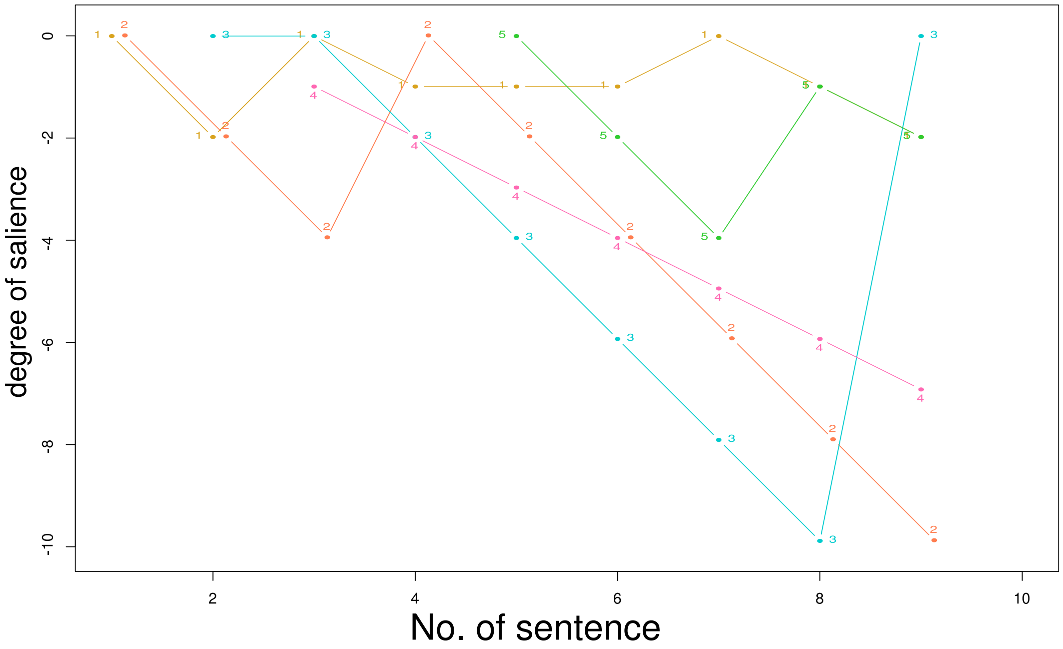

Figure 4.7: Example of a short document from PDiT along with its salience graph.

(1) Accounter and one million have disappeared

(2) Brno

(3) Since 11th June, when (he) left the work around 3 AM and did not come home, the police is searching for a 27-year-old Stefan Misik, main accounter of casino 777 on the Svobody square in Brno.

(4) The searched-for man had over million crowns with him and could be a victim of a violent crime.

(5) Stefan Misik resides in Pradlacka street and has a well-built, 178-cm-high figure, short brown hair and a pea-sized birthmark on a left side of his neck.

(6) During the speech, (he) burrs.

(7) Last time (he) was wearing (on him) a bright shirt, black jeans and brown loafers.

(8) On the neck, (he) was wearing a silver chainlet with a sign of Cancer, in a black bag had also a new passport and magnetophone tapes.

(9) Witnesses can report to the nearest police office, the 158 (phone) line or the I. department of Crime Service in Brno, phone 05/4116 2525.

4.2.2 Salience Graph Generation Procedure

One of the main parts of this work was to automatize the procedures needed for the visualization of the salience for each document. This consists of several steps, the whole process being summarized in 4.2.2. Each step is performed by a procedure in a script file, making the intermediate results analyzable. Each of these script files uses a language which seemed the most appropriate for the task: When working with the PML files, `btred` is used, `Perl` itself is employed for non-PML text manipulations, and `R` language was chosen for the graph visualization part. Some parts of the scripts were originally created during the preparation of Hajičová et al. (2006). However, their ad-hoc nature made them largely impossible to suit our purposes, thus they were all significantly rewritten, made more readable, documented, and hopefully reusable.

The first step is to modify the PML files by identifying the nodes of each coreference string and marking them accordingly – this process is often called "coloring the coreference strings". This is achieved by applying a simple algorithm of linearly going through the tectogrammatic tree nodes, inspecting their direct coreference antecedents and denoting them by the according color number identifier. The next small step, rather technical, is to order the color identifiers sequentially with respect to the linear flow of the sentences (this process actually is not necessary for the functionality, rather a convenience for further inspection).

Computing the salience degree of each coreference string members is done in the subsequent step. This is where the salience algorithm is applied on the colored nodes. In each sentence in the "colored" PML files, salience degree is computed for each coreference string which has appeared so far, and extracted into an external file. This information, serving as "coordinates", is then fed into the R script described further.

The actual graphical form of the salience graphs is generated by a script in the R programming language. As its input, it is given a set of files (each corresponding to one document) with salience "coordinates": for each occurence of a coreference string member, there is a line in the file with its coreference string identifier, sentence number and the salience degree of the member's occurence. From this coordinates of each point of the salience graph, a graphical file is generated – either as a bitmap (PNG) or in a vector-based format (SVG or PostScript). The output is made as readable as possible, providing both colors and numbers for each coreference string curve, as well as a slight shifting of the curves to reduce their overlaps. However, the variability of the salience behavior of the strings, inherent density of the curves in a large part of the documents and the variability of the documents' lengths make it hard to effectively generalize some of the techniques

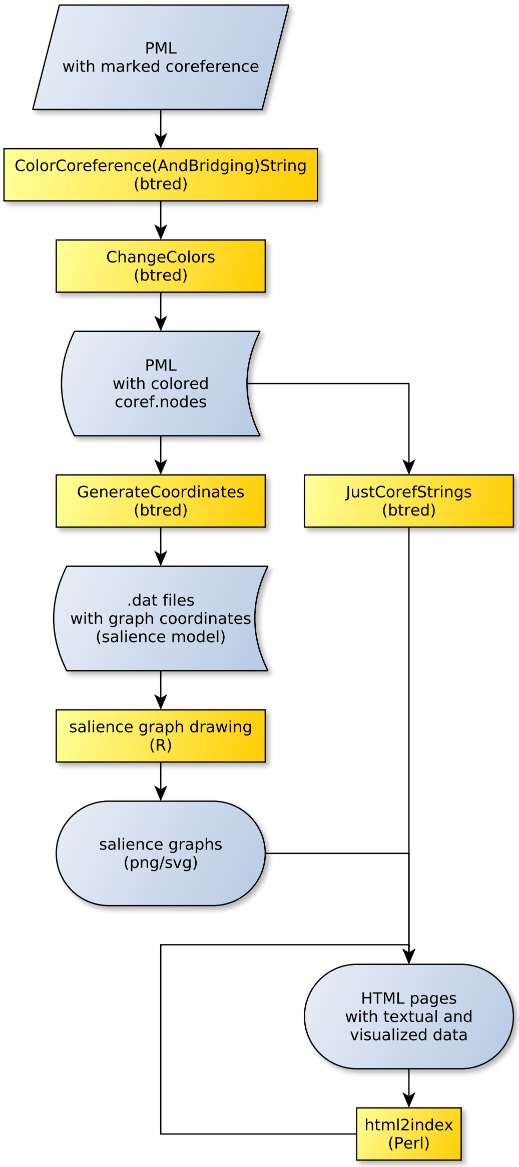

Figure 4.8: Flowchart of the data processing from PML corpus data to the salience graph visualization embedded in an HTML page.

used for improving the readability and clarity of the graphs.

4.2.3 Vertical Cut

Moving on the “horizontal” axis, i.e. sentence by sentence, and observing the current trend of all the chains at once, certain vertical breaks can be identified in the salience models. These suggest a slight change of topic in the particular sentence, where several new objects emerge or re-activate and the old ones fade away. From this point of view, the salience models can be used e.g. for automatic segmentation of previously unsegmented text by “cutting” the text at this breaks, perhaps into paragraphs. Furthermore, the objects emerging at the identified breaks (or later in the beginning segment) can suggest the topic of the current segment. The design of an actual algorithm for such automatic process is not covered by this work, although one should be able to test it rather conveniently on the PDiT data with the original paragraph segmentation preserved, thus applicable as the gold standard.

4.2.4 Horizontal Cut and Leap Height

Another approach to the models would be to draw one or more horizontal lines in the graph to mark a certain level of salience. One can assume that these levels can express the amount of activation of an object must have to be referred to by certain grammatical means – a weak or zero pronoun is expected to refer to an object with high activation, whereas less salient objects are re-activated by more specific expressions, e.g. a definite noun phrase.

To verify these hypotheses, let us introduce a new quantity: *salience leap height*, or simply *leap height*. Each time an object (represented by its coreference chain, i.e. chain of expressions referring to it) is mentioned in a sentence, the *leap height* value indicates the difference of its current salience level and its salience level in the previous sentence. More rigorously, let the leap height value of an object x (or, from another point of view, of its coreferents' chain) in sentence number n (where x is mentioned) be defined as such:

$$LeapHeight(x,n) := dg_n^x - dg_{n-1}^x \tag{4.1}$$

Note that this definition contains not only the “depth” from which the mentioned object emerges, but takes into account also the TFA value of the current referring expression, in the form of its current salience value – being it either 0 or -1. This reflects the idea of differentiating the referent's actual sentence

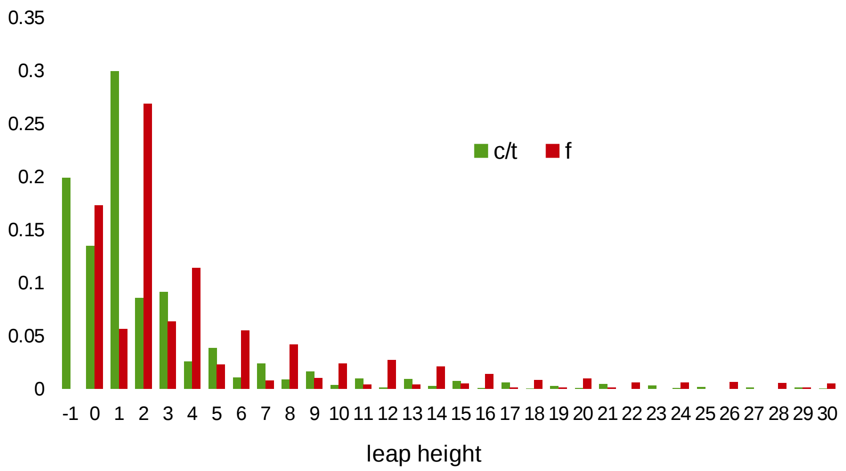

Figure 4.9: Proportions of the leap heights comparing the coreferents' TFA values; from *train-1* data. (The y-axis units are ratios of leap heights for the given *sempos* value normalized to sum to 1.)

function. This differentiating is proportionally more important with the smaller leap heights and losing its importance with their higher values, which may not necessarily be harmful. This property also results in a possibility of the leap height having a zero value, or even a negative value, specifically -1; when the last reference of x was in the focus (had TFA value f) of the previous sentence and the current reference is in the topic (has TFA value t or c). This situation is actually quite common in the discourse; it corresponds to the usual case of a newly emerged object in the $(n-1)$-th sentence, which is subsequenly referred to in the n-th sentence, serving in it as a "starting point" (a topic, in the TFA terms).

All the leap-heght charts presented in this section has their values normalized to sum up to 1 within the given feature value. The reason is that in these analyses, we are mostly interested on the distribution within the given value, rather than directly comparing the two absolute values at any fixed leap height.

Leap Heights and TFA Figure 4.2.4 shows the frequency of the leap heights depending on the TFA value of the referring expression. A general rule may be stated that shorter leaps are typical for mentioning in topic (c/t), while the longer ones are slightly more common for mentioning in focus (f).

Also note the fact that the leaps to the topic are apparently more frequent for the odd leap heights, whereas the focus "destination" favors the even leap heights. This is an inherent property stemming from the inclusion of the TFA in the definition of the leap height.

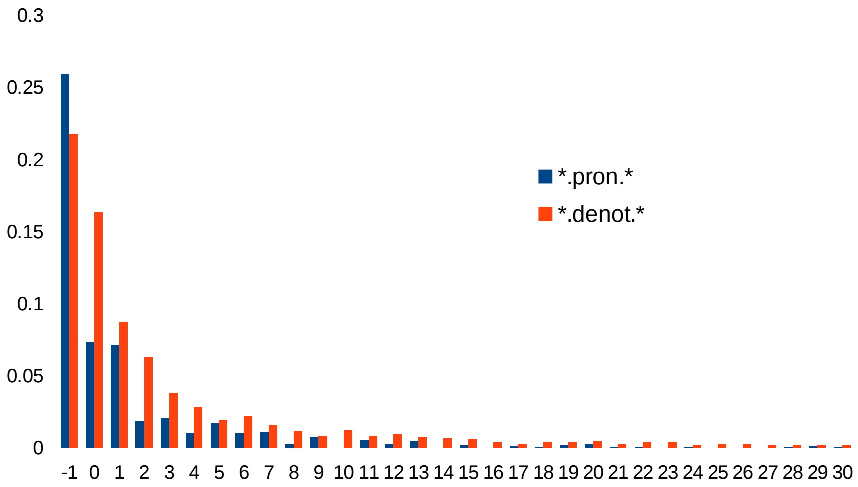

Figure 4.10: Proportions of the leap heights for the chosen *sempos* categories; from *train-1* data. (The y-axis units are ratios of leap heights for the given category normalized to sum to 1.)

Pronominal vs. denominating referents Let us return to the above mentioned hypothesis about the grammatical form of referents typical for certain salience ranges. Thanks to an elaborate system of the tectogrammatical layer annotation in PD(i)T, we can use the t-node attribute *sempos*[2]. The pronominal expressions are marked with *sempos* value containing *.pron.* (e.g. $n.pron.indef$ standing for "indefinite pronominal semantic noun"), whereas the *sempos* value of the denominating expressions contains *.denot.* (e.g. *n.denot* means "denominating semantic noun"); the rest being only quantificational expressions and verbs. With this division, we can visualize the proportions of the leap heights within each of these *sempos* categories in Figure 4.2.4[3].

From the chart, it is obvious that there is some disproportion in the behavior of the pronominal referents in comparison to the denominating ones. The quick drop of the pronominals' values beyond the leap height of 1, along with the rather steady decline of the denominators, seems to confirm the declared hypothesis. However, the dominance of the -1 value is quite surprising and calls for a deeper analysis. The Figure 4.2.4 thus focuses only on comparing demonstrative and personal pronouns (*sempos* values $n.pron.def.demon$ and $n.pron.def.pers$, respectively), because these two are by far the most frequent types among the pronominal coreferents. The difference between them is apparent: while the demonstrative pronouns almost fails to refer beyond the leap height of 1 and

[2] From the PDT t-layer annotation manual: "The sempos attribute (semantic part of speech) contains the information regarding the membership of a complex node in a semantic part of speech." (Hajič et al., 2006)

[3] Although the leap height values goes as far as 172, the tail is long and its values neglectable for our purposes – thus the charts are often cut off at the leap height value of 30

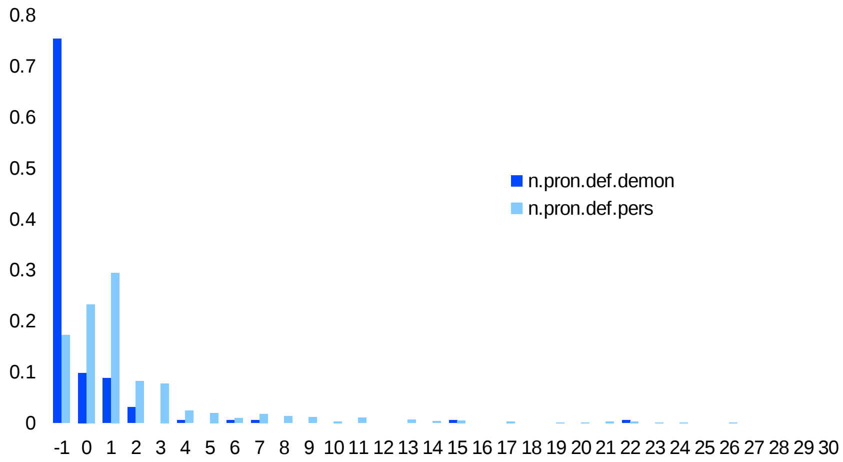

Figure 4.11: Proportions of the leap heights for the chosen two pronominal *sempos* values; from *train-1* data. (The y-axis units are ratios of leap heights for the given *sempos* value normalized to sum to 1.)

serves mostly for the −1-leap reference, the personal pronouns, although also "specialized" on the low leaps, perform best for the leaps of 1 or 0. From this comparison, it is also evident that the demonstrative pronouns were almost fully responsible for the high values of leap height −1 for pronominals in the previous categorial comparison.

5. Topic Modeling Experiments

5.1 Preliminary Experiment – Clustering

Before performing the experiments with evaluating a topic modeling application, one preliminary experiment was conducted to form an idea of how the salience information could contribute to the document information. A visual comparison of the document relations was created between the document information contrasting simple word counts against average salience of coreference chains.

This visualization was based on representing the document collection as a graph – with documents as the graph nodes and pairwise document *overlap* defining the graph edges. The contrastive comparison was then made by changing the definition of how the *overlap* is computed. The generic idea was for each document to list some of its characteristic items in the order of their supposed importance, cut this list off at some point, and then look for matching items in the other documents' lists.

5.1.1 Sorting the Nouns by Counts

One of the most straightforward and yet most frequently used features of extracting key words from a document is the word count. Usually it is complemented by a filter of stop-words, but in our case, when we have the information about the word types at our disposal, the simplest way is to work with nouns only.

5.1.2 Sorting the Chains by Average Salience

Average salience, adjusted When looking for an optimal measure for ordering the whole coreference chains in terms of a coarse informative representativeness, the average salience is a natural choice. However, to avoid favoring chains which first occur lately in the document, their salience must be adjusted to better reflect their inactivity before their first occurence. According to the idea that these chains are in the stock of shared knowledge, but not mentioned, their initial course is simulated similarly as if they had been mentioned in the topic of the first sentence. Thus, until their first mention, they undergo a descent by 1 from the value of -1. The general formula for computing the average salience of chain referring to an object x is then as follows:

$$\begin{aligned} AvgSal(x) :&= \frac{1}{N}\left(\sum_{i=1}^{m-1}(-i) + \sum_{i=m}^{N} dg_i^x\right) \\ &= \frac{1}{N}\left(-\frac{(m-1)m}{2} + \sum_{i=m}^{N} dg_i^x\right) \end{aligned} \tag{5.1}$$

where m is index of the sentence with the first mention of x and N is the total number of sentences in the document.

Having defined the average salience, the graph of document overlaps based on that measure can be constructed analogously:

For each document:

1. for each coreference chain, extract the list of nouns from the chain along with the average salience of the whole chain,
2. sort the chains according to their average salience,
3. cut the list at 10% of its length, so that the most salient chains remain.

Then, the *overlap* of two documents is defined as the number of overlapping chain pairs from the lists of these documents.

5.1.3 Clustering Visualization

A collection of documents can be viewed as a graph where each node represents one document. Two documents have a common edge iff there is a (non-zero) *overlap* between them, and the weight of this edge equals the size of this overlap. Then, with the help of a commonly used graph visualizing tool (in our case, pygraphviz, see Section 3.3.3), the two graphs resulting from the definitions above were drawn for a visual evaluation only. Having no "correct" results of a real proximity or topic relationships, any numeric evaluation would be hard and was omitted here because of the preliminary nature of the experiment. However, the goal of this experiment was to get a basic idea of how the salience information could alter the results of a computational analysis of the document collection. The resulting graph visualization of the noun-based overlaps and salience-based overlaps can be seen on Figures 5.1 and 5.2, respectively.

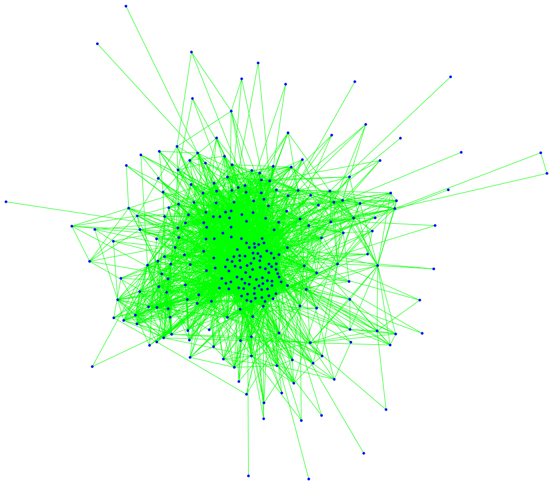

Figure 5.1: Noun-based document overlap.

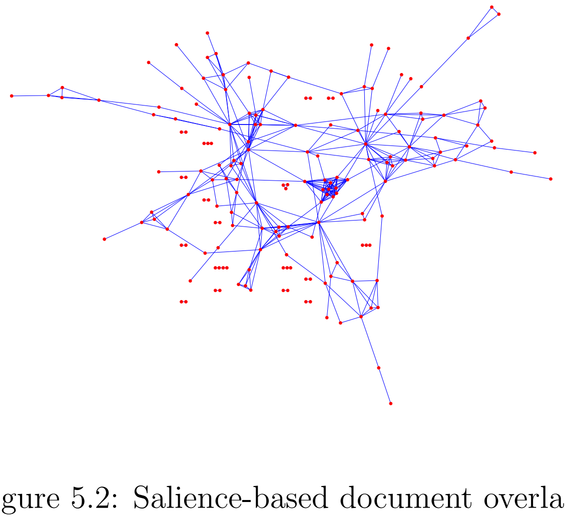

Figure 5.2: Salience-based document overlap.

5.2 Latent Dirichlet Allocation and Salience

The definition of the salience values can be also viewed as assigning some momentary importance to each object mentioned in the document and represented by a set of its co-refering expressions. And the notion of importance is actually a key feature in the discipline of information retrieval, where given a set of documents, one is trying to find the most important and distinguishing keywords, or topics. It is natural to assume that knowing the salience values would help in this pursuit.

During the preliminary research of the LDA approach and for the first experiments with it, Python-based `gensim` framework was used. This was primarily because of its customization options, modularity and easy access and possibility of modification of its data structures. However, due to an absence of a reasonable interface for an automatic evaluation, `gensim` was eventually abandoned in favor of Java-based `MALLET`. This wide-ranged NLP processing toolkit widely used for the topic modeling tasks provides a robust usage of the LDA computation. It is not so directly customizable as `gensim`, but still modular to a certain extent, and offers a straightforward way for evaluating the LDA results automatically, which was a key factor in this decision.

Among the possible LDA evaluation techniques, we have chosen the use the left-to-right algorithm, based on the arguments presented in (Wallach et al., 2009) (see also Section 2.4.1). Concerning the speed of the algoritm applied in our settings: on a common laptop hardware configuration, both the training on *train-1* (2533) and the evaluation of 150 topics on the *dtest* (316 documents) usually lasted not much more than one minute.[1]

5.2.1 LDA without Salience

Before we will try to employ the salience into the LDA machine learning algorithm, we have to establish a baseline for the upcoming comparison. Thus, we have performed a series of experiments of applying the LDA on the (preprocessed) PDiT data and evaluating it using the left-to-right procedure.[2]

During these experiments, only lemmata of nouns were used. This was a result of two directions of reasoning, besides a simplicity and easy human assessment of

[1] The processing time of the two phases should be considered together, because lots of preparation and pre-calculations for the evaluation phase is already being done in the training phase, during the model building – hence the existence of the “evaluator” file created in the training phase.

[2] For the theoretical details, see Section 2.4.1.

the results. The first reason relates to a common practice in an environment of bag-of-words models: establishing a list of stopwords and leaving them out during a preprocessing phase. This is motivated by their very little informativeness across a set of documents. Our idea was taken from the opposite-side view – keeping only the supposedly most informative set of words. When the linguistic annotation is at hand, the most straightforward approach was to assume that the nouns are the most informative class (in the sense of their key-ness). The second reason concerned the upcoming comparison and anticipating the form of engagement of the salience information. Since the approach was going to be representing each coreference chain with one of its members, the most comparable variant seemed to be again the restriction to nouns only. Indeed, from this point of view, this is definitely not a well designed universal baseline; if another approach was be chosen for the salience involvement, the results would hardly be comparable.

Gridsearch for parameters Since the LDA is a model parametrized by several variables, a grid-search was performed on *train-1* to find the most promising pair of at least two most important for them; the number of topics and the main α parameter. Subsequently, usually only this pair was used in the further experiments, sometimes supported by another one or two more promising pairs to confirm the results.

5.2.2 LDA with Salience

Coercing chains to words The main design decision in this phase was how to transform the salience-related data to the document format needed as an input for the LDA computation. We are actually looking for a way to convert the rich information about the annotated documents in PDiT, with emphasis on the coreference chains and their salience values, to a simple bag-of-words document model to feed into the LDA. Although this is probably the most determining decision in this experiment, we have eventually settled for a quite simple idea:

> *The document will consist of a bag of representants of coreference chains only. Each coreference chain will be represented by the lemma of its first member, and this lemma will be repeated n-times in the document, where n is an adapted value of the average salience*[3] *of the coreference chain.*

[3] For the average salience definition, see Equation 5.1.

Let us present some notes as a reasoning behind this definition:

- First member of a chain as its representant – this was led by the natural intuition (supported by the data) that the first mentioning of the referred object is almost always the most specific. Thus, when comparing chains across documents, this representant should function as a quite efficient identificator of the referred objects, matching when the objects are the same and vice versa.

- Multiplicity of representants – this notion exploits the bag-of-words document represantation of LDA which discards any information about word ordering and keeps just the distinct word counts.

- Average salience – similarly to the preliminary clustering experiment (Section 5.1), we are working with coreference chains as a whole and in the bag-of-words model, the only quantitative information expected is their frequency count. Thus, the average salience comes as a natural choice of the only one number to be "disguised" as the count, carrying the information about the chain importance within the document.

The "adaptation" of the average salience mentioned in the definition is a solution to the problem of converting the negative decimal value (average salience) to an integer value (word count) while preserving its monotonicity. For this task, the following simple numeric conversion was devised:

$$AvgSal(x)' := \left\lceil -\frac{100}{AvgSal(x)} \right\rceil \tag{5.2}$$

There is an obvious side-effect of this definition following its non-proportional nature: the chains with an extremely low value of the average salience will have a very similar resulting "word-count" (usually 1 or 2). This might be intuitively beneficial, since at the extreme poles of a scale, even large differences are percieved as relatively small.

5.3 Performance measures

The main performance measure of the comparison experiment is the output of the left-to-right evaluation method from the trained model. Its output values are logarithms of probabilities of heldout documents given the trained topic distributions. In our case, the topic distribution models were trained either on *train-1*, or on *train-all* dataset, each preprocessed according to the definitions described in Section 5.2. The *dtest* dataset was used as the held-out data for all the experiments performed, preprocessed accordingly.

The logarithms of probabilities (or simply log-probabilities) are monotonous to the original probabilities, so the inter-comparability is preserved in this sense.

5.4 Results

In the Tables 5.1 through 5.4, the log-probabilities of the above described experiments are displayed and summarized. The discussion of these results follows.

α topics	0.1	0.5	1	5	10	20	50
10	-228956	-227344	-226574	-226294	-226541	–	–
20	-226801	-223922	-223414	-222086	-222251	–	–
50	-225177	-220072	-218890	-217602	-217292	-217513	-218919
100	-226753	-219534	-217944	-215475	-215399	-215326	–
150	-226677	-220368	-218126	-215543	**-214672**	-215179	–
200	-227646	-221380	-218760	-215787	-215547	–	–
500	-231848	-226808	-223626	-219749	-218982	–	–

Table 5.1: Results of grid-evaluation of num-topics and α parameters; log-likelihoods trained on *train-1* and evaluated on *dtest* noun-lemmata.

α topics	10	20
100	-208491	-209504
150	**-207766**	-208547

Table 5.2: Results of the best main parameter pairs with training on *train-all* and evaluated on *dtest* noun-lemmata.

5.4.1 Discussion

Unfortunately, the results presented in the Tables 5.1-5.4 does not suggest at all that the salience information used in the way described in the previous sections should bring any improvements to the LDA results. However, the radical differences in the results suggest rather that there is a substantial flaw in the whole experiment settings.

Especially the results in Tables 5.3 and 5.4, which seemingly favors models with far more topics than there are documents in the held-out dataset (or in *train-1*) points to a fundamentally wrong design of the document synthetization.

α topics	0.1	0.5	1	5	10	20
10	-1291347	-1284963	-1280990	-1277357	-1275835	–
20	-1237530	-1221743	-1216872	-1213659	-1211691	–
50	-1165614	-1137140	-1129198	-1121563	-1116290	–
100	-1111397	-1068746	-1059458	-1046104	-1042701	–
150	–	–	–	–	-995740	–
200	-1025492	-1006016	-956812	-969008	-906184	–
500	-987313	-930654	-915353	-869368	**-857240**	–

Table 5.3: Results of grid-evaluation of num-topics and α parameters; log-likelihoods trained on *train-1* and evaluated on *dtest* with average salience-word counts.

α topics	10	50
150	-861420	-1002466
200	-830238	-970996
500	-738437	**-847686**

Table 5.4: Results of the best main parameter pairs with training on *train-all* and evaluated on *dtest* with average salience-word counts.

With this kind of results at hand, it is impossible to confirm or disprove the possibility of the benefits of salience for an NLP application like topic modeling. Evident is that some crucial decisions about the experiment design in this chapter were wrong and any possible future research should try to avoid those directions.

Conclusion

We have presented a reproduction and a data-oriented analysis of the salience algorithm formulated earlier, along with visualizing its results and confirming some of the hypotheses behind the salience notion. This was achieved using the data of the Prague Discourse Treebank 1.0, especially its annotation of the coreference relations and the topic-focus articulation. A brief experiment with the bridging anaphora annotation data was conducted in an attempt for broadening the coverage of the salience models, but deeper investigation in this field remains to further research.

The visualization procedure suggested earlier was made more robust and automatized to allow larger amount of documents to be processed. Also it was extended with procedures which makes the results human-accessible even in this scale.

Another key features of this work were the attempts to interpret the output of the salience procedure, the salience graphs. A notion of *salience leaps* and their height was introduced and used to confirm the hypothesis about the importance of salience in the decisions about the syntactic form of the referent.

Finally, two series of experiments in the area of document processing were performed to estimate a possible contribution of the salience information in this field. However, the results especially of the second one were unfortunately unsatisfactory, due to its poorly designed settings.

Let us hope that at least the first part of the work stimulates a further research in this undoubtedly promising area, while the second part might serve as an indication of an impasse direction.

Bibliography

Barzilay, R. and Lapata, M. (2008). Modeling Local Coherence: An Entity-Based Approach. In *Computational Linguistics*, vol. 34(1):pp. 1–34.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. In *The Journal of Machine Learning Research*, vol. 3:pp. 993–1022. ISSN 1532-4435.

Chafe, W. L. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In *Subject and Topic* (edited by C. N. Li), pp. 25–55. Academic Press, Cambridge, MA, USA.

Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A Framework for Modeling the Local Coherence of Discourse. In *Computational Linguistics*, vol. 21(2):pp. 203–225. ISSN 0891-2017.

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., and Mikulová, M. (2006). *Prague Dependency Treebank 2.0.*

Hajičová, E. (1993). *Issues of sentence structure and discourse patterns.* Charles University Press, Prague, Czech Republic.

Hajičová, E. (2003). Contextual boundness and discourse patterns. In *Proceedings of XVII International Congress of Linguists, CD-ROM*, pp. x1–x7. Matfyzpress, MFF UK, Prague, Czech Republic.

Hajičová, E. (2013). Contextual boundness and discourse patterns revisited. In , vol. 15(5):pp. 535–550.

Hajičová, E., Havelka, J., and Veselá, K. (2005). Corpus Evidence of Contextual Boundness and Focus. In *Proceedings of the Corpus Linguistics Conference Series*, pp. 1–9. University of Birmingham, Birmingham, UK.

Hajičová, E., Hladká, B., and Kučová, L. (2006). An Annotated Corpus as a Test Bed for Discourse Structure Analysis. In *Proceedings of the Workshop on Constraints in Discourse Structure Analysis*, pp. 82–89. National University of Ireland.

Hajičová, E. and Vrbová, J. (1982). On the role of the hierarchy of activation in the process of natural language understanding. In *Proceedings of the 9th conference on Computational linguistics*, vol. 1 of *COLING '82*, pp. 107–113. Academia Praha, Prague, Czechoslovakia.

Halliday, M. A. K. (1967). Notes on Transitivity and Theme in English: Part 1. In , vol. 3(1):pp. 37–81.

Joshi, A. K. and Weinstein, S. (1981). Control of Inference: Role of Some Aspects of Discourse Structure-Centering. In *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 385–387.

Kuno, S. (1972). Functional Sentence Perspective: A Case Study from Japanese and English. In , vol. 3(3):pp. 269–320.

Kučová, L. and Hajičová, E. (2004). Coreferential Relations in the Prague Dependency Treebank. In *Proceedings of the 5th International Conference on Discourse Anaphora and Anaphor Resolution Colloquium*, pp. 97–102.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit.

Minka, T. and Lafferty, J. (2002). Expectation-propagation for the Generative Aspect Model. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, pp. 352–359. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. ISBN 1-55860-897-4.

Nedoluzhko, A. (2011). *Rozšířená textová koreference a asociační anafora. Koncepce anotace českých dat v Pražském závislostním korpusu.* ÚFAL – Institute of Formal and Applied Linguistics, Prague, Czech Republic.

Prince, E. F. (1981). Toward a taxonomy of given-new information. In *Syntax and semantics* (edited by P. Cole), vol. 14, pp. 223–255. Academic Press, New York.

Sauper, C., Haghighi, A., and Barzilay, R. (2010). Incorporating Content Structure into Text Analysis Applications. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pp. 377–387. Association for Computational Linguistics, Stroudsburg, PA, USA.

Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects.* Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands.

Susan E. Haviland, H. H. C. (1974). What's new? Acquiring New information as a process in comprehension. In , vol. 13:pp. 512–521.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50. ELRA, Valletta, Malta. `http://is.muni.cz/publication/884893/en`.

Wallach, H. M. (2008). *Structured topic models for language.* Ph.D. thesis, University of Cambridge.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation Methods for Topic Models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 1105–1112. ACM, New York, USA. ISBN 978-1-60558-516-1.

List of Figures

2.1 Graphical model representation of the unsmoothed LDA model in a plate notation (source: (Blei et al., 2003)). 11

3.1 Example of coreference annotation for the following sentence: *Pro zásobování Ostravska a Frýdeckomístecka potřebuje firma svá jatka. (The company needs its slaughterhouse in order to supply the Ostrava and Frydek-Mistek regions.)* The dark red arrow is used for a grammar coreference relation, dark-blue arrows for textual coreference; light-blue arrow for bridging reference (to an expression in another sentence). 16

4.1 Distribution of the per-document sentence counts in *train-1* dataset. 20

4.2 Distribution of the per-document sentence counts in *train-all* dataset. 20

4.3 Counts of the coreference link types in *train-1* and *train-all* dataset. 21

4.4 Frequency of lengths of coreference chains in *train-1* dataset; cut off at length of 26 nodes. 22

4.5 Frequency of lengths of coreference chains in *train-1* dataset – the impact of adding bridging anaphora. 23

4.6 Frequency of TFA values in *train-1* dataset. 24

4.7 Example of a short document from PDiT along with its salience graph. 25

4.8 Flowchart of the data processing from PML corpus data to the salience graph visualization embedded in an HTML page. 27

4.9 Proportions of the leap heights comparing the coreferents' TFA values; from *train-1* data. (The y-axis units are ratios of leap heights for the given *sempos* value normalized to sum to 1.) . . . 29

4.10 Proportions of the leap heights for the chosen *sempos* categories; from *train-1* data. (The y-axis units are ratios of leap heights for the given category normalized to sum to 1.) 30

4.11 Proportions of the leap heights for the chosen two pronominal *sempos* values; from *train-1* data. (The y-axis units are ratios of leap heights for the given *sempos* value normalized to sum to 1.) . . . 31

5.1 Noun-based document overlap. 34

5.2 Salience-based document overlap. 34

List of Tables

4.1 General statistics of the datasets. 19

4.2 Coreference type link counts 21

4.3 Counts of the whole coreference chains in the datasets, related to numbers of documents. 21

5.1 Results of grid-evaluation of num-topics and α parameters; log-likelihoods trained on *train-1* and evaluated on *dtest* noun-lemmata. 38

5.2 Results of the best main parameter pairs with training on *train-all* and evaluated on *dtest* noun-lemmata. 38

5.3 Results of grid-evaluation of num-topics and α parameters; log-likelihoods trained on *train-1* and evaluated on *dtest* with average salience-word counts. 39

5.4 Results of the best main parameter pairs with training on *train-all* and evaluated on *dtest* with average salience-word counts. 39

Appendix – CD-ROM Contents

- `data` folder – Contains sample data from the PDiT 1.0 corpus (`pdit_sample`). All the files with the `.t.gz` extension in this folder are in the PML format and ready to be used by the `btred scripts` (however, some of the `btred` scripts will modify them, so make sure the data files have a write permission and you have a backup of them before running those scripts).

- `scripts` folder – All the non-trivial script files used in this work; `btred`, `Perl`, `Python`, `R` files, `bash` scripts. Most of them require to be run on a Linux machine, `btred` scripts require the `btred` application to be installed. The `scripts-readme.txt` file provides the overview of the script files along with a brief information about their functionality and usage.

- `vacl-dipl_thesis.pdf` file – This work in pdf format.

- `readme.txt` file – General information about contents of the CD-ROM.