MAGISTERARBEIT

SEMANTIC DESCRIPTION OF TEXT CLUSTERS

Ausgeführt am Institut für
Software Technology and Interactive Systems
der Technischen Universität Wien
unter der Anleitung von ao.univ.Prof. Dr. Andreas Rauber
durch

Nataliya Sokolovska
Premreinergasse 19/2/5, 1130 Wien

Wien, am 31. Dezember 2005
Abstract

To process natural language using biologically inspired algorithms of neural computation is one of the challenges nowadays. The great amount of information needs new, fast and efficient algorithms and methods that take not only single words into consideration, but the context of a word as well.

Using Self-Organizing Maps, it is possible to process data efficiently, but it is still a question, what could be done for a better interpretation of the SOMs.

In this thesis we present methods and tools that compare areas of the map according to several methods. Each unit of a SOM has several characterizing words, which are divided into four layers of specificity. We consider LabelSOM labels to be the first and most general layer, keywords represent the second layer, GATE annotations and context words of GATE annotations are the third and fourth layers respectively.

The context of words which belong to the same semantic category is examined. The words in this context are weighted and annotated with part-of-speech tags. In order to give more efficient weights to the context words, we differentiate between the word instance context and the type context of a semantic category. We propose a new weight for context words that is based on the keyword selection algorithm of K. Lagus and S. Kaski.

We demonstrate the potential of the developed tools in the field of information retrieval, using standard text collections in different languages and of different size for our experiments.
## Contents

1 Introduction .................................................. 8

2 SOM-based text clustering .................................. 12
   2.1 Text indexing, weighting, vector normalization, and stoplists ........ 12
   2.2 Self-Organizing Maps ................................... 14
      2.2.1 Structure ........................................ 14
      2.2.2 Initialization of the map .......................... 15
      2.2.3 Learning process ................................. 16
   2.3 Summary .................................................. 17

3 Semantic labels, keywords and annotations ............... 18
   3.1 Keyword selection method for characterizing text document maps .... 18
      3.1.1 Keywords for map clusters ........................ 19
      3.1.2 Keywords for map units ............................ 20
   3.2 LabelSOM method ....................................... 21
      3.2.1 Algorithm of getting labels ....................... 22
   3.3 GATE ..................................................... 23
      3.3.1 ANNIE and language plugins ....................... 23
   3.4 Summary .................................................. 25
4 From text to self-explaining maps of information

4.1 Text collections ........................................... 26
4.2 Preprocessing and indexing ................................. 28
4.3 Somtoolbox: training maps ................................. 29
4.4 Using GATE .................................................. 30
   4.4.1 Corpus creation and annotation ....................... 30
   4.4.2 XML structures of annotations, type and instance contexts of annotations .......... 31
   4.4.3 Weights of context words and the best context .......... 33
   4.4.4 Comparative summary of two areas ...................... 36
4.5 Technical description of tools ............................. 37
   4.5.1 Tool “Comparison of map areas” ...................... 37
   4.5.2 Tool “Semantic description of a map area” .............. 38
   4.5.3 Graphical user interface .............................. 41
4.6 Summary .................................................. 42

5 Experiments ................................................ 44

5.1 Collection Time Magazine ................................. 44
   5.1.1 Information provided for a unit ....................... 44
   5.1.2 Brief description of the map ......................... 46
   5.1.3 LabelSOM labels versa keywords ..................... 50
   5.1.4 Comparison of two units ............................. 51
   5.1.5 Comparison of two clusters .......................... 55
   5.1.6 Comparison of a unit and a cluster ................... 58
   5.1.7 Investigation of a map area ......................... 60
5.2 Collection Help Government (help.gv.at) .................. 63
5.2.1 Brief description of the map ................................. 63
5.2.2 Comparison of two units .................................. 68
5.2.3 Comparison of two clusters ............................... 71
5.2.4 Comparison of a unit and a cluster ....................... 75
5.2.5 Investigation of a map area ............................... 75
5.3 BankSearch DataSet 7.0 ....................................... 77
  5.3.1 Brief description of the map ............................. 77
  5.3.2 Comparison of two units ................................. 80
  5.3.3 Comparison of two clusters ............................. 81
  5.3.4 Comparison of a unit and a cluster .................... 81
5.4 Summary .......................................................... 82

6 Conclusions ......................................................... 83

A Notations .......................................................... 86
List of Figures

1.1 SOM (BankSearchDataSet 7 text collection. D-Matrix) .......................... 9
1.2 Selecting a cluster. ................................................................. 9

2.1 SOM with hexagonal and rectangular grids. ................................. 15
2.2 Adaption of weights of a winner and its neighbors. ....................... 16

3.1 ANNIE components. ............................................................... 24

4.1 Scheme of text processing. ...................................................... 27
4.2 Semantics of text clusters. Main menu. ..................................... 27

5.1 Time Magazine collection. (D-Matrix) ....................................... 46
5.2 LabelSOM labels and keywords. .............................................. 52
5.3 Time magazine collection. Comparison of two units. ................. 52
5.4 Time Magazine collection. Comparison of two units. Semantic zooming. 54
5.5 Time Magazine collection. Comparison of two clusters. ............... 56
5.6 Time Magazine collection. Comparison of two clusters. LabelSOM method. 57
5.7 Time Magazine collection. Comparison of clusters using GATE annotations. 59
5.8 Time Magazine collection. Investigation of a map area. Weighted context. 61
5.9 Time Magazine collection. Comparative summary of two clusters. ...... 62
5.10 HelpGov collection. (D-Matrix) ............................................... 63
5.11 HelpGov collection. Comparison of two units. ........................................ 70
5.12 HelpGov collection. Comparison of two units. Semantic zooming. ........ 70
5.13 HelpGov collection. Comparison of clusters. ........................................... 73
5.14 HelpGov collection. Comparison of clusters according to keywords. ..... 74
5.15 HelpGov collection. Comparison of clusters according to GATE annota-
    tions. ........................................................................................................... 74
5.16 HelpGov collection. Comparison of clusters according to context of GATE
    annotations. ............................................................................................... 74
5.17 Time Magazine collection. Comparison of clusters according to context
    of GATE annotations. ................................................................. 74
5.18 HelpGov collection. Annotations found in the area. .............................. 76
5.19 HelpGov collection. Weighted annotations of the selected files. .......... 77
5.20 HelpGov collection. Finding the best context. ................................. 78
List of Tables

4.1 Analysis of context words. 35

5.1 Lexical items provided for a unit. 47

5.2 Comparison of two clusters based on GATE annotations. 72
To my Granny Irina
and my Brother Sergey
Chapter 1

Introduction

The great amount of information demands new techniques and tools. The question how to process text fast and efficiently, not manually, is an important research challenge.

Self-Organizing Maps (SOMs) are one of the neurally inspired algorithms. They are used for different tasks, among them text clustering. The important distinction of SOMs is cleanness of visualisation. If input data are present in a complicated structure, it is difficult and not always possible to find some correlation between elements. Using the SOM [15], it is possible to simplify a multidimensional structure. Hence, SOMs can be considered to be one of the methods that project a multidimensional space to a space with fewer dimensions. Usually it is a two-dimensional map, which can be simply shown on a computer display.

Another distinction of this model is its learning method. Self-Organizing Maps follow an unsupervised learning process. They are able to reflect a data structure, and that is why they are a good choice for clustering tasks.

Using SOMs, we can cluster text according to e.g. its content. We get maps, that visualize the similarities between documents in terms of distances within a two dimensional map. Hence, similar documents may be found in neighboring regions of the map
Figure 1.1: SOM (BankSearchDataSet 7 text collection. D-Matrix)

<table>
<thead>
<tr>
<th></th>
<th>49</th>
<th>4</th>
<th>16</th>
<th>7</th>
<th>28</th>
<th>38</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1.2: Selecting a cluster.
display.

On one hand, it is quite easy to interpret the map visualization: documents are arranged according to common words, on a basis of which a common topic can be imagined. Documents, that correspond to the same or similar topic, are situated close to each other, e.g. in the same unit, in neighboring units or in the same cluster. On the other hand, it is not easy to determine which attributes are relevant for a mapping.

We need to search for much more, in our case, textual information, that characterizes units or areas of the map. Documents that occur in a unit, must be investigated more thoroughly.

While processing text in a natural language, one is not allowed to forget the context factor. Lexical items, which are keywords chosen by an algorithm, can have a quite different linguistic context and can be different notions.

A natural question, which arises while looking at a SOM, is whether it is possible to obtain more information about a unit, or a cluster. In this work we consider textual information. Our aim is to review existing methods and algorithms, to analyze the possibilities to provide units with more semantic information and to investigate some selected area of a map. Another question is what is common and different between two units, two clusters, a unit and a cluster. The question is particularly interesting, if two areas are neighbors on a map.

Figure 1.1 represents a trained SOM of size $25 \times 35$. One can see a number of units. Each of units contains information about a number of documents that were put into it and labels (obtained using the LabelSOM method, which we discuss later). The labels help to give a semantic interpretation for a unit and brief generalization for all documents that are put into a unit. Figure 1.2 shows the possibility to select several units of a map and to create a cluster.
This work is organized in the following way. In Chapter 2 we review the principles of text clustering, the SOMs algorithm and the notions that are connected to that: text indexing, weighting, vector normalization and stoplists.

Chapter 3 consists of three main parts. We describe the keyword selection method for characterizing text document maps, the LabelSOM method and a tool that was developed for natural language processing, GATE.

Chapter 4 is dedicated to the preprocessing steps, such as creating template and vector files, a map training and getting GATE annotations. In this Chapter we give a technical description of the developed tools as well. A brief description of Java classes and their functionality is provided.

Chapter 5 describes experiments. Three text collections – Time Magazine collection, HelpGovernment web sites and BankSearchDataSet 7 collection – of different size (420, 1024 and 11000 documents) and in different languages (English and German) are taken. Comparisons of a unit with a unit, of a unit with a cluster, of a cluster with a cluster and the investigation of a map area are carried out for two smaller collections. Every map unit is provided with labels on different layers of precision. We consider LabelSOM labels to be the first and most general layer. Keywords are the second layer, GATE annotations and context words of GATE annotations represent the third and fourth layers respectively. As a final step of this work, we create a comparative summary of two areas of the map.

The largest collection BankSearchDataSet 7 is examined according to fewer criteria because of a high complexity provoked by its size.

In Chapter 6 we demonstrate results of experiments and provide conclusions and ideas for future work.

The Appendix A gives a list of notations that we use in our thesis.
Chapter 2

SOM-based text clustering

In this chapter, we review the problem of text clustering, notions that are connected with this problem and an algorithm to cluster a text collection. Using SOM [14], it is possible to arrange a number of text documents in such a way that similar documents are grouped close to each other. We examine the SOM rather formally and in detail; we describe, how it is initialized and how the network learns.

2.1 Text indexing, weighting, vector normalization, and stoplists

The notions of a document and of a single word are important in Information Retrieval. Each document of a collection is represented by a vector of \( n \) weighted index terms. The values of weights belong to the interval \([0, 1]\). The weighting scheme \( tf \times idf \) (term frequency times inverse document frequency) is introduced to measure the importance of a word for representing the content of a document.

A weight \( w_{jk} \) is defined for each lexical item \( t_k \) of a document \( d_j \).
The standard tfidf weighting scheme is defined by G. Salton and C. Buckley [22] as following:

\[ tfidf(t_k, d_j) = \frac{\#(t_k, d_j) \cdot \log \frac{|C|}{\#(t_k)}}{\#(t_k)} \]  

(2.1)

where \( \#(t_k, d_j) \) is the number of times \( t_k \) occurs in \( d_j \), \( \#(t_k) \) is the number of documents in \( C \) in which \( t_k \) occurs at least one time. \( C \) is the training set.

The meaning of this formula is the following: a word is considered to be important for a document, if it occurs frequently in this document and not often in others.

Weights have to be in the interval \([0, 1]\). The vectors, that represent the documents, have to be of the same length. That is why the weights are normalized by cosine normalization:

\[ w_{jk} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{\#T} (tfidf(t_s, d_j))^2}}, \]  

(2.2)

where \( T \) represents all the terms that occur at least once in \( C \).

While processing a corpus that includes several (sometimes hundred or more) documents, one gets a large number of terms. The complexity of the task of text processing is still too high for existing algorithms and hardware. But one can see, that not all of the terms are important and some of them could be deleted without losses. There are two possibilities of dimension reduction, each of which includes several methods:

1. Reduction by feature selection.

- The most intuitive way to reduce by feature selection is to remove words that are in a stoplist. These are words that belong to the functional category of a lexicon and that do not possess semantic information, e.g. articles, conjunctions, pronouns, and auxiliary verbs. The drawback of this method is that it is language dependent and for every language such a stoplist must be created.
• A method using the document frequency \( \#(t_k) \). If a term appears only in few documents, it is possible to suppose that this term is not important. Terms that are present in a large number of documents or which occur only in few could be removed, because they do not provide a sufficient distinction between documents in an input space.

• A method using the term occurrence frequency \( \#(t_k, C) \). The value defines, how many times a term appears in the whole collection. If a term occurs very often or very seldom, it could be removed.

2. Reduction by feature extraction.

• Stemming is a morphological method, in which a word is reduced to its root form. Thus, different inflections of the same word are mapped to one single, common root.

• Using a thesaurus, it is possible to substitute specific words with a more general notion.

2.2 Self-Organizing Maps

2.2.1 Structure

Self-Organizing Maps [14] are one of the algorithms of neural computation. It is an unsupervised tool, hence, the result of learning depends only on the structure of the input data. SOMs define a mapping from the input data space \( \mathbb{R}^n \) onto a two-dimensional array of nodes. The mathematical model includes a layer of input and output units. Input units are responsible for providing output units with input patterns. With every node \( i \) (output unit), a parametric reference vector \( m_i = [\mu_{i1}, \mu_{i2}, ..., \mu_{in}]^T \in \mathbb{R}^n \) is
associated. Output units can be arranged in a rectangular, hexagonal, or even irregular grid (see Figure 2.1). In the simplest case, an input vector \( x = [\xi_1, \xi_2, ..., \xi_n]^T \in \mathbb{R}^n \) is connected to all neurons (output units) in parallel via variable scalar weights \( \mu_{ij} \), which are in general different for different neurons. In an abstract scheme it may be imagined that the input \( x \), by means of some parallel computing mechanisms, is compared with all the \( m_i \), and a location of the best match in some metric is defined as the location of the “response” [15].

### 2.2.2 Initialization of the map

A configuration of a grid (rectangular or hexagonal) and a number of neurons are given in advance. It is necessary to initialize weight vectors of neurons before learning. Usually one uses one of two methods to initialize weight vectors:

1. Initialization with small random values;

2. Initialization using samples (some random samples from a training set are taken as initial values).
2.2.3 Learning process

The learning process consists in a sequence of corrections of output weight vectors. One of the input vectors is chosen and is compared to all neuron weight vectors on every learning step. The most similar neuron to an input vector is called the winner. Under similarity of an input vector and an output unit is meant the smallest Euclidean distance between them. Let us examine Figure 2.2. We see an adaption of weights of a winner and its neighbors on some learning step. Coordinates of an input vector are shown with a cross, coordinates of nodes after adaption are shown with a black color and a grid itself after modification is shown with strokes.

More formally, let \( x \in \mathbb{R}^n \) be a data vector. One might then say that the SOM is a “nonlinear projection” of the probability density function \( p(x) \) of the high-dimensional input data vector \( x \) onto a two-dimensional display. The vector \( x \) is compared to all \( m_i \), and the smallest of the Euclidian distances \( ||x - m_i|| \) defines the best-matching node, signified by \( c \):

\[
c : ||x - m_c|| = \min_i ||x - m_i||
\]  

(2.3)
For the adaption of weight vectors the following equation is used:

$$m_i(t + 1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)], \quad (2.4)$$

where $t = 0, 1, 2, \ldots$ is a current training iteration and $x(t)$ is a random vector from the input.

The function $h(t)$ is a neighborhood function, which depends on the time and the distance between the winner and its neighbors.

$$h_{ci} = \alpha(t) \cdot \exp \left( -\frac{||r_c - r_i||^2}{2\sigma^2(t)} \right), \quad (2.5)$$

where $\alpha(t)$ is a scalar-valued “learning-rate factor”, and the parameter $\sigma(t)$ defines the width of the kernel. Both $\alpha(t)$ and $\sigma(t)$ are monotonically decreasing functions of time.

### 2.3 Summary

In this chapter we have reviewed the task of text clustering and the SOM algorithm. If the input data is high dimensional, and if the goal is to project them onto a space with fewer dimensions, e.g. onto a two-dimensional space in order to visualize the input data, then the task can be solved using the SOMs. The task of text clustering is exactly this case, because text is represented in a high dimensional input space.
Chapter 3

Semantic labels, keywords and annotations

3.1 Keyword selection method for characterizing text document maps

In this chapter we deal with the question, what the SOMs learn and which common features are responsible for arranging some documents into the same cluster and some documents on the opposite sides of a display. The aim of the method of K. Lagus and S. Kaski [16] is to characterize units, clusters, and map areas containing several similar documents. A keyword selection method is proposed that can be used for obtaining descriptions of data clusters. Some methods that cluster the data sets provide an order or distance measure of the clusters. K. Lagus and S. Kaski suppose, that if such an ordering of the clusters exists or can be deduced, then it is possible to develop a method that uses this order and provides keywords for units and clusters.

If the input data is textual, then the descriptions which needed to be found are
keywords that characterize the collection of text documents of a particular area [16].
These keywords, placed on the map, help a user to find the necessary information and describe areas of the map.

The method is especially suitable for characterizing maps of data collections organized by the SOM algorithm. The method focuses on the distributions of words occurring within the document groups. And it can be used in order to put labels for a group of similar documents.

Although it were more logical to start with the keywords for units of a map, we consider keywords for clusters first. The reason is, that K. Lagus and S. Kaski introduce their method in such a way, that the algorithm of finding unit labels is based on the cluster labelling.

### 3.1.1 Keywords for map clusters

K. Lagus and S. Kaski [16] give the following definition of a descriptor or of a keyword:

A good descriptor characterizes some outstanding property of a cluster in relation to the rest of the collection. That is, if a word $t_k$ is a good keyword for a cluster $U$, then $t_k$ has the following properties:

1. $t_k$ is prominent in $C$ compared to other words in $C$.

2. $t_k$ is prominent in $C$ compared to the occurrence of $t_k$ in the whole collection.

These criteria are combined to a general measure for the purpose of keyword ranking:

$$G(t_k) = F^{\text{clust}}(t_k) \cdot F^{\text{coll}}(t_k), \quad (3.1)$$

where $F^{\text{clust}}$ characterizes a word $t_k$ in relation to other words within the cluster $l$, the term $F^{\text{coll}}$ relates the word $t_k$ to the whole collection.
Let $f_\ell(t_k)$ be the number of times the word $t_k$ occurs in the cluster $\ell$, i.e. the frequency of the word $t_k$ in $\ell$. Then let $F_\ell(t_k)$ be the relative frequency of a word $t_k$, defined as

$$F_\ell(t_k) = \frac{f_\ell(t_k)}{\sum_v f_\ell(v)} \quad (3.2)$$

It should be noted that $0 < F_\ell(t_k) < 1$ and $\sum_{t_k} F_\ell(t_k) = 1$. Such a normalization is done to disregard the size of a cluster and to measure the relative importance of a word compared to the other words occurring in the cluster. The relative frequency $F_\ell(t_k)$ describes the relation between the term $t_k$ and other words within the cluster $\ell$, and that is why the relative frequency $F_\ell(t_k)$ is a good candidate for $F_\ell^{clust}$. So, one can write:

$$F_\ell^{clust} = F_\ell(t_k) \quad (3.3)$$

$F_\ell^{coll}$ measures the relation of the frequency of $t_k$ in the cluster $\ell$ to the “background frequency” that describes how typical the word is within the whole collection. Without doubts, a straightforward measure for this value is:

$$F_\ell^{coll} = \frac{F_\ell(t_k)}{\sum_p F_p(t_k)} \quad (3.4)$$

where $\ell$ and $p$ are the cluster indices. The component $F_\ell^{clust}$ helps to find words that take up a large proportion in their cluster. $F_\ell^{coll}$ is responsible to find words that are dominant within the cluster with respect to the whole document collection.

So, the goodness $G$ of a word $t_k$ appearing in cluster $\ell$ is defined as:

$$G(t_k, \ell) = F_\ell(t_k) \frac{F_\ell(t_k)}{\sum_p F_p(t_k)} \quad (3.5)$$

### 3.1.2 Keywords for map units

The approach of finding keywords for map units is based on an assumption that units have an order, i.e. the neighbors of a map unit are more similar to it than the units far
away.

So, if a word $t_k$ occurs often in a map unit $\ell$, a word is probably also frequent in some adjoining area of the map. K. Lagus and S. Kaski decided to exclude an area of the map immediately surrounding a unit $\ell$ in the calculation of the “inhibitory factor” $F_{\ell}^{\text{coll}}(t_k)$, so:

$$G(t_k, \ell) = F_{\ell}(t_k) \frac{F_{\ell}(t_k)}{F_{\ell}(t_k) + \sum_{p \in B_{\ell}^1} F_p(t_k)},$$  \hspace{1cm} (3.6)

where $p \in B_{\ell}^1$ if $d(p, \ell) < r_1$; $d(p, \ell)$ is the distance on map grid between units $p$ and $\ell$. $B_{\ell}^1$ is a “neutral map zone” around the unit $\ell$. This neutral zone is excluded from determining the goodness of a word $t_k$ as a keyword for a unit $\ell$.

## 3.2 LabelSOM method

LabelSOM defines the elements of a vector (input data) that are the best match to characterize a unit. The most appropriate descriptors of a unit are attributes that are shared by all data of a unit.

So, it is necessary to determine those vector elements from each weight vector which, on one hand, exhibit about the same value for all input signals mapped onto that specific node as well as, on the other hand, have a high overall value indicating its importance.[20]

The quantization error vector $q_i$ is used in order to measure the variance of every vector element.

It is computed for every unit $i$ and it is an accumulated distance between the weight vector elements of all input signals $x_j$ mapped onto unit $i$ and the unit’s weight vector elements.

$$q_{ik} = \frac{1}{|C_i|} \sum_{x_j \in C_i} |(m_{ik} - x_{jk})|$$  \hspace{1cm} (3.7)

21
where $k = 1 \ldots n$. Here, $C_i$ is the set of input patterns $x_j \in \mathbb{R}^n$ mapped onto a unit $i$. Summing up the distances for each vector element $k$ over all the vectors $x_j$ ($x_j \in C_i$) yields a quantization error vector $q_i$ for every unit $i$.

Having got a quantization error vector $q_i$, it is possible to define for every attribute $k$, what is the probability that all attributes are characteristic for all data of a unit. The smaller is $q_i$, the more probable is that an attribute is shared by all data to the same degree.

However, there is a disadvantage of this method. The method provides attributes that are either present or not present on a unit to the same degree for a unit. The algorithm functions perfectly, if one wants to find attributes that are shared by all data, but in case of LabelSOM another opportunity to find attributes that are not present, is not interesting. Thus, we try to find attributes, that are shared by all data. The algorithm [21] is described below.

### 3.2.1 Algorithm of getting labels

Let SOM have $i = x \times y$ units, and $m_i$ be weight unit vectors of dimension $n$. The following algorithm is carried out for every unit of a map.

1. Quantization error vector is calculated according to (3.7);

2. A list of candidates for labels is created. In order to choose labels from this list it is possible to use some $\lambda$ candidates with the lowest quantization error vectors or to use a threshold $\tau_1$ and select those labels, for which $q_i < \tau_1$;

3. From the list of candidates select those with weight vector value $m_i > \tau_2$. 

22
3.3 GATE

GATE 3.0 [8] is a software architecture for language engineering.

GATE possesses an original software architecture. The elements of GATE can be divided into three different types, that are called resources:

1. Language Resources include lexicon, corpora, or ontologies;
2. Processing Resources include algorithmic elements, e.g. parsers, generators;
3. Visual Resources are visualization and editing components of the GUI.

3.3.1 ANNIE and language plugins

A family of Processing Resources that is integrated within GATE and that is of special interest for this work, is called ANNIE, A Nearly-New Information Extraction system. ANNIE is oriented on the English language and consists of several components, that subsequently process a text in natural language.

ANNIE includes components to process tokens (universal tokens) and English tokens; ANNIE is able to split the sentences, to annotate a text with part-of-speech tags and semantic tags. The ANNIE components are shown on the Figure 3.1.

The universal tokenizer splits a text into primitive tokens, such as punctuation, numbers, and words. The English tokenizer’s task is to find abbreviations and shortened forms of lexical items that are characteristic for the English language. Gazetteers are plain text files that include lists of annotations, such as names of organizations, cities, months, etc. The sentence splitter of ANNIE is responsible to annotate every sentence of a text with a type “sentence”. The sentence splitter is domain independent. The part-of-speech tagger provides every word of a document with a part-of-speech tag, and the semantic tagger produces sets of annotations. The semantic tagger is based on the
Figure 3.1: ANNIE components.
JAPE Grammar (or JAPE language, Jave Annotation Patterns Engine) [3] that defines annotation rules.

An annotation is a form of meta-data attached to a particular section of the document content [8], it is a lexical item and is defined by the content it comes from, by its start and end points in the content. Every annotation has a type, that is defined by a JAPE language. Gazetteers define values that annotations can take.

The language plugins for several languages are distributed with the last version of GATE (January, 2005, version 3.0). It is possible to process texts in French, German, Spanish, Chinese, Arabic, Romanian, Hindi and Cebuano.

In our work we use GATE to annotate document collections, used in our experiments. As a result of GATE processing, we get sets of lexical items, that refer to the same semantic categories, for example, a set of words that refer to the notion “location”.

3.4 Summary

In this chapter we discussed the LabelSOM method and the keyword selection method to describe the SOMs. In this work we use both of these methods to characterize map units.

GATE is a tool for natural language processing. Nowadays, it supports not only English, but a number of other languages. What is especially important for our work is that, with the help of GATE, one obtains annotations from a text – groups of words that are semantically similar. The categories of annotations are slightly different for different languages.
Chapter 4

From text to self-explaining maps of information

In this chapter, we describe the whole way (see Figure 4.1) that is to be gone from the “raw” text to the SOMs which are trained on a text collection, and to the map interpretation.

4.1 Text collections

We have three different text collections. They are:

1. Time Magazine collection: news from the 1960ies. The collection is in English and consists of 420 documents.

2. HelpGovernment collection: web sites from help.gv.at. There are 1,026 html documents in the text collection. Most of documents are in German, but about 5% of the articles are not in German.
Figure 4.1: Scheme of text processing.

Figure 4.2: Semantics of text clusters. Main menu.
3. BankSearchDataSet 7.0: BankSearchDataSet 7 is in English. The collection is provided by the University of Reading. It consists of 11,000 text documents.

We process each collection separately, but we use the same methods for all three collections.

4.2 Preprocessing and indexing

In Chapter 2 we explained the concepts of text indexing, weighting, vector normalization and stoplists. In order to create vectors, we use the program TeSet [19]. The program provides different possibilities to reduce the number of terms. An option “min_length” allows to define the minimal length of a term, “max_length” - maximal length of a term, “min_df” is a minimal document frequency, and “max_df” is a maximal document frequency. For all our collections we use “min_length” = 3, and “max_length” = 30, that means that terms shorter than three symbols and longer than thirty letters will be deleted. As to other two important parameters, namely “max_df” and “min_df”, we pay attention to the collection size. For every text collection a corresponding list of stop words is used.

For the Time Magazine text collection we get:

- initial setup: documents: 420, terms: 20,829;
- reduction: df_min 6 and max_df 300... remaining 3,399 terms;
- english stopword list ... remaining 3,113 terms.

For the HelpGov collection:

- initial setup: documents: 1,026, terms: 35,415;
• reduction: df_min 10 and max_df 400... remaining 3,274 terms;

• german stopword list ... remaining 3,082 terms.

In the experiment with the BankSearchDataSet collection we get:

• initial setup: documents: 11,000, terms: 306,622;

• reduction: df_min 75 and max_df 500... remaining 5,091 terms;

• english stopword list ... remaining 4,953 terms.

4.3 Somtoolbox: training maps

For clustering and visualization we use a tool called Somtoolbox. Somtoolbox is used in two steps: training of a map and visualization of results of the SOMs training. For the first step we need a template file and a vector file that we produce with the help of TeSet. We define the dimensions of a map which depend on the size of a text collection. For the Time Magazine collection, that has 420 text pieces and it is our smallest collection, we train a map with $10 \times 15$ units. The HelpGovernment collection consists of 1,026 documents, and we decide to train a map $15 \times 25$. Our third collection, BankSearchDataSet is much bigger than the previous two, and we select a map $25 \times 35$. In Chapter 2 we reviewed, how SOMs learn, and gave definitions for the notions of learning rate. In our case the learning rate is 0.7 and the number of iterations equals 100,000. We discuss the second step, visualization of results and their interpretation, in the next chapters.
4.4 Using GATE

4.4.1 Corpus creation and annotation

There exists an implemented infrastructure for the English language in GATE (ANNIE, already considered above). As to the German language, there is a newly developed plugin (included in the last GATE version [8], described above as well). The ANNIE modules, discussed above, are the basis for processing of all languages, although the gazetteer lists are language specific.

We use GATE to get semantically similar groups of lexical items – words or phrases – annotations. It is possible to use GATE both as a GUI tool, and as a library. We use GATE as a library. Our goal is to use this library in order to create our Java class(es) that process the documents.

Under a corpus we understand a set of documents. GATE creates an empty corpus, to which somehow marked documents can be added.

A GATE application is performed in the following way:

1. The processing resources of ANNIE or a language plugin are loaded. If ANNIE is used, then the following processing resources are loaded:

   - gate.creole.annotdelete.AnnotationDeletePR,
   - gate.creole.tokeniser.DefaultTokeniser,
   - gate.creole.gazetteer.DefaultGazetteer,
   - gate.creole.splitter.SentenceSplitter,
   - gate.creole.POSTagger,
   - gate.creole.ANNIETransducer,
   - gate.creole.orthomatcher.OrthoMatcher;
2. A corpus is created and populated with documents to be processed;

3. Documents are searched for annotations. If an annotation is found, the start and end positions are saved. These points of content define an annotation;

4. Found annotations are sorted according to their positions in the content;

5. Annotations are ready for further processing. In this work for example, we get the context of annotations and annotations themselves.

The GATE annotations for the English language are “Person”, “Location”, “Organization”, “Address”, “Age”, “Offence”, “Drug”, “Time”, “Date”, “Percent”. We use the default set of annotations. The GATE application for the German language has a different set of annotations. The annotation set that we use to get the annotations is called NE and contains the following annotations: date (“DATE”), exception (“Exception”), location (“LOC”), number (“NMB”), organization (“ORG”), person (“PER”), time (“TIME”), profession (“PROF”), percent (“PER”).

There is a number of annotations in the same annotation set, that do not have semantic information (“token”, “split”, “spacetoken”, “sentence”, “default_token”) and which therefore do not possess enough useful information.

4.4.2 XML structures of annotations, type and instance contexts of annotations

The aim of this work is not only to annotate the text collections. The goal is to get and process annotations and the context of annotations in such a way that it were possible to use the result of this processing for SOMs description.

We want to create some specific structures and to process every text collection in three different ways. We get a new folder of XML files, derived from a folder of an
original text collection. Each file of the new folder has the same name as in the original folder. Each XML file is divided into three parts:

1. Annotations. There is no meaningful text anymore, there are only pure annotations, e.g.

   <ORG>Plenum</ORG>
   <DATE>2004</DATE>

   In this example, “ORG” is a type, and “Plenum” is a value.

2. Pieces of text that we consider to be a context (we call it type context, i.e. context of specific annotation type) of an annotation. This XML structure is created in the following way: an annotation is found and its type is defined. A string “context” is added to the type. The context is defined as a text fragment that surrounds an annotation and that is limited from left and from right by one of the delimiters “,”, “.”, “?”, “;”. For example,

   <ORG/context>
   November 2004 wurde das Pensionsharmonisierungsgesetz
   im Plenum des Nationalrates beschlossen
   </ORG/context>

   <DATE/context>
   November 2004 wurde das Pensionsharmonisierungsgesetz
   im Plenum des Nationalrates beschlossen
   </DATE/context>

3. Pieces of text that we consider to be a context as well (we call it instance context of an annotation). This XML structure is created in a similar way to the type
context. The difference is that in this case we get an annotation itself instead of a type of an annotation. In case of annotations that are numbers or annotations that contain numbers, we take all the letters of an annotation to the “-context” string and add an “id”, a value of which consists of all numbers found in an annotation.

4.4.3 Weights of context words and the best context

The GATE annotations are weighted according to the above described method of K. Lagus and S. Kaski [16]. Equation (3.5) is used to weight annotations.

The context of an annotation consists of several words in natural language. It was already mentioned, there are two kinds of context: the type context and the instance one. Both contexts can be used to weight context words. We suppose that the best weight for a context word is the combination of the type context weight value and the instance context weight value.

In this work we propose a new weight for the context words, that is based on the keywords selection method, described above.

Let \( G_{\text{inst}} \) be the goodness of a context word counted within the instance context, and \( G_{\text{type}} \) be the goodness of a context word counted within the type context for a specific
type. We propose the following relation, that shows the importance of a context word:

\[ G_{context} = \frac{G_{inst}}{G_{type}} \]  

(4.1)

The smaller the value of \( G_{context} \) is, the more important is a word in a context. In other words, the more specific a word is, the more characteristic it is for a unit.

Let us consider an example. The process of SOM training puts two files of Time Magazine collections, \( T190.html \) and \( T426.html \) into the same unit (9/12). The unit is about Congo politics and its prime minister Adoula. The word “adoula” is selected by the LabelSOM method and is a keywords of this unit too. The method of GATE annotations annotates “adoula” as a person.

Let us see, how context words are selected. “adoula” is found in the following contexts: “adoula’s government has been on the razor’s edge of defeat at the hands of leopoldville’s chaotic legislature”, “said adoula”, and “this is just one of many problems of leopoldville’s premier adoula”. The English stop list is used and a number of words are deleted. Words that remain are examined (see Table 4.1). They are put into the first column of the table. The second and the third columns demonstrate instance context values (\( G_{inst} \)) and type context values (\( G_{type} \)). The next column provides \( G_{context} \). We see, that words “premier”, “defeat”, and “government” have the smallest \( G_{context} \) values. At the same time, giving a look at the fifth column, presenting keywords values, we see, that keyword method provides the following words: “government”, “hands”, “defeat”, “premier”. If we compare words, given by \( G_{context} \) value and keywords, we come to the conclusion, that the sets of words are almost identical and the word “hands”, which is put as the second according to its weight, is not so appropriate for this unit. Hence, “premier”, “defeat”, and “government” are selected as the best context words for “adoula”.

If we have weighted words of each context and we can tell, which word is more
<table>
<thead>
<tr>
<th>Context words</th>
<th>Instance context value</th>
<th>Type context value</th>
<th>Instance context value/Type context value</th>
<th>Keyword value</th>
</tr>
</thead>
<tbody>
<tr>
<td>razor’s</td>
<td>1.0405827263267429E-4</td>
<td>1.314924391847468E-5</td>
<td>0.01263</td>
<td>0.8025102520684703E-5</td>
</tr>
<tr>
<td>chaotic</td>
<td>1.0501995379122034E-4</td>
<td>1.035143108534755E-5</td>
<td>0.09856</td>
<td>0.4585772868962687E-5</td>
</tr>
<tr>
<td>leopoldville’s</td>
<td>20.811654526534857E-4</td>
<td>1.4792899408284025E-5</td>
<td>0.00710</td>
<td>0.40125512603423E-5</td>
</tr>
<tr>
<td>edge</td>
<td>1.0405827263267429E-4</td>
<td>0.6574621959247344E-5</td>
<td>0.00631</td>
<td>0.080251025206847E-5</td>
</tr>
<tr>
<td>legislature</td>
<td>1.0405827263267429E-4</td>
<td>0.3287310979618672E-5</td>
<td>0.00316</td>
<td>0.072955477460770E-5</td>
</tr>
<tr>
<td>hands</td>
<td>1.0405827263267429E-4</td>
<td>0.3114294612270321E-5</td>
<td>0.00299</td>
<td>3.210041008273881E-5</td>
</tr>
<tr>
<td>defeat</td>
<td>1.0405827263267429E-4</td>
<td>0.14610271020527432E-5</td>
<td>0.00140</td>
<td>2.0062756301711757E-5</td>
</tr>
<tr>
<td>premier</td>
<td>5.202913631633714E-4</td>
<td>0.05689576695493856E-5</td>
<td>0.00109</td>
<td>1.805648067154058E-5</td>
</tr>
<tr>
<td>problems</td>
<td>1.4863258026159333E-4</td>
<td>0.05428743131961345E-5</td>
<td>0.00365</td>
<td>0.114643217240671E-5</td>
</tr>
<tr>
<td>government</td>
<td>1.0405827263267429E-4</td>
<td>0.017822770371426536E-5</td>
<td>0.00017</td>
<td>3.210041008273881E-5</td>
</tr>
</tbody>
</table>

Table 4.1: Analysis of context words.
important than the others, we can make an assumption, that it is possible to use these weights in order to find the best context. We sum up the weights of context words and divide this value by the number of words in this context.

### 4.4.4 Comparative summary of two areas

The comparative summary is produced using the LabelSOM labels, keywords, GATE annotations, and best context of the GATE annotations. The summary provides common and different LabelSOM labels and keywords. The set of common words and one of different are divided into three groups according to their weights. It is done in the following way. The highest and the lowest weights are found, the difference between the values is calculated and the result is divided by 3. So, we get three groups of words: ones with the high weights (are marked in the summary with the red color), ones with medium (blue), and ones with low weights (green). The second part of the summary uses the GATE annotations and their best context. GATE annotations which have the highest weight in a semantic category, are marked with the bold font and their best context is given in brackets.

Summary is based on a primitive template that consists of the following components:

- “According to the LabelSOM method both areas deal with”
- “Area 1 focuses on (LabelSOM method)... and area 2 on ”
- “Area 1 focuses on (Keywords)... and area 2 on ”
- “Both areas tell about the same common locations: ”
- “Both areas are about the same personalities: ”
- “In these two areas one can get information about the following organizations: ”
4.5 Technical description of tools

In this section we describe two tools that allow a user to investigate an area of a map and to compare areas of a map according to lexical items (LabelSOM labels, keywords, GATE annotations and context words of GATE annotations). We provide a description of the Java classes and of a graphical user interface.

4.5.1 Tool “Comparison of map areas”

Requirements for the tool

The tool is integrated into the Somtoolbox. The tool “Comparison of map areas” is carried out using NetBeans 4.1, Java Development Kit 5.0 and requires the latest version of JDK (5.0) to perform all its functions. The tool was successfully tested on the Windows XP platform and on Mandrake 10.0.

Input and output of the tool

The input of the program consists of two important parts:

1. the output files of the TeSet program - vector and template files, and a property file - the standard input for the Somtoolbox;

2. a folder of files that are XML structured and that are created by a Java program that uses GATE functions. We developed a Java program for this purpose, that is language specific.
The result of the comparison of two units, two clusters, or of a unit and a cluster are lists of common and different words that occur in these areas.

It is possible as well to see LabelSOM labels, keywords, GATE annotations and context words of GATE annotations for every selected unit.

Java classes and their functions for this tool and classes of the tool “Semantic description of a map area” will be described together in the next section.

4.5.2 Tool “Semantic description of a map area”

Requirements for the tool

The requirements are the same as for the tool “Comparison of map units”.

Input and output of the tool

The input of the tool is identical to the input of the already described tool “Comparison of map areas”.

The output of the tool includes:

1. all the annotations of the given categories
2. the weighted annotations of the given categories
3. a context - piece of text - of each annotation (the context is empty if an annotation is within a html tag and if the semantic information of context is not important)
4. all context weighted words of an annotation with part-of-speech tags

Java classes and their functions

In this section we describe the Java classes of the tool and the graphical user interface. We provide some screenshots in order to give a visual impression.
AnnotationProcessing

The class AnnotationProcessing is responsible for the output of annotations of the selected area according to the categories. There are the following semantic categories in GATE [4] for the German language: “Person”, “Location”, “Time”, “Organization”, “Profession”, “Exception”, “Percent”, “Date”, “Number”.

CompareByFrequency

The class CompareByFrequency is responsible for ordering words from the least frequent to the most frequent. Words are put in an alphabetical order if their frequency is equal.

KeywordAnnotationProcessing

The class KeywordAnnotationProcessing is responsible for the output of keywords and their goodness (calculated with the help of the method of K.Lagus and S. Kaski [16]). The categories of the annotations are the same as for the AnnotationProcessing class.

ReadAndInputXML

The class ReadAndInputXML is created in order to give a possibility to read XML files and to get the arrays of annotations from them. As an input the class gets the LinkedList of the files that are necessary to process.

ReadAndInputXMLString

The class ReadAndInputXMLString is created in order to give a possibility to read XML files and to get the arrays of annotations from them. As an input the class gets the String of the files that are necessary to process.

SemanticSOMPanel

The class SemanticSOMPanel is responsible, first of all, for graphical menu representation. The menu possesses seven options:
• to provide the annotations of a semantic category;

• to provide the annotations of a semantic category with its weight;

• to show the context of annotations, to weight each word of a context, and to add a part-of-speech tag for each context word;

• to compare selected areas of a map according to the LabelSOM method;

• to compare selected areas of a map according to keywords method;

• to make a comparison of selected areas according to GATE annotations;

• to make a comparison of selected areas according to the context of GATE annotations;

• to provide the comparative summary of selected areas.

**WordFrequency**

The class WordFrequency provides the frequency for every word of selected files. The class possesses special functions that allow to process a frequency of a word, whether the words are in a String or in a LinkedList. In this class there is a useful method “ignoreFile()” that allows to enter a stop-words-file into the command line.

**ContextMain**

The class ContextMain is responsible for the most complex option: it shows the weighted annotations of a category, and for every annotation it provides its context, a weight for every word in a context, and a part-of-speech tag for every context word. A context is empty if an annotation is within a html tag and if semantic information of the context is not important.

**LabelSOMComparison**
The class LabelSOMComparison is responsible for visualizing the results of the comparison of LabelSOM labels and keywords.

**KaskiGateComparison**

The class KaskiGateComparison is responsible for visualizing the results of the comparison of GATE annotations of two areas of a map or of context words of GATE annotations.

**Summary**

The class Summary provides the comparative summary of selected areas. The summary is based on a primitive template. The parameters of the comparison are the LabelSOM labels, keywords, GATE annotations, and context words of the GATE annotations. The LabelSOM labels and keywords are divided into three parts according to their importance. The most important words, i.e. those, which possess the highest weights, are marked with the red color. Words with the lowest weights are green and keywords, that occur in the category between these two, are blue. The most important GATE annotations are bold.

### 4.5.3 Graphical user interface

We tried to make the graphical user interface as friendly and intuitive as possible. The graphical user interface is built into the Somtoolbox tool as a separate menu “Semantics of text clusters”. The menu has six options and a screen to show the files of selected areas (Figure 4.2).

A user selects two areas on the map (they will be shown in two corresponding fields). Then it is possible to perform a comparison of the areas. If the button “LabelSOM method” is pressed, a user will see different and common words of two areas according to the LabelSOM method. If the button “Keywords” is pressed, it is possible to see
different and common words of two areas according to the Keywords method. The output of “Keywords + GATE” is the comparison based on GATE annotations of two units (not all of GATE annotations are taken into consideration, but only those, that possess the best goodness according to the Keywords method). If a user presses the last button “Keywords + GATE + Context”, one gets the comparison of two areas according to the context of GATE annotations. The keywords of these four categories are sorted according to their weights.

If the button “Annotations of the selected area” is pressed, a user gets all the annotations put into categories that are found in this area. The button “Keywords & Goodness of Annotations” shows the annotations with their weights. So, a user can consider which annotations could be descriptors for the area. The button “Annotations with Context” opens a window, in which a user can select a category. According to the selected category the annotations with their weights are shown. A user can select an annotation and get all weighted lexical items of all contexts and sorted contexts.

The button “Comparative Summary” provides a summary that is close to a natural language. It takes common and different LabelSOM labels and keywords, common and different GATE annotations. The most important GATE annotations are given with its best context and are marked with the bold font. LabelSOM labels and keywords are presented in red, blue, and green colors. The red words possess the highest weights, blue – the medium ones, and green – the lowest weights.

4.6 Summary

The creation of units consists of several important steps and is carried out with the help of the tools TeSet and Somtoolbox. GATE is used to annotate the documents of a collection. We described our own approach to the weighting of context words in
this chapter. Special roles are played by the XML structures, created using GATE as a library, the notion of word instance and type contexts, and the method of finding keywords for the map units.

In this chapter we described the tools, provided brief description of Java classes and their functions. The tools have a friendly GUI and are easy to use. The only important thing a user has to know before running it, is how to interpret the weights of the context words: the smaller a value of a weight is, the more important is a context word. The context here is a complex notion, it includes an instance context and a type context according to the semantic category.
Chapter 5

Experiments

5.1 Collection Time Magazine

The collection Time Magazine includes 420 text documents which are abstracts of news from the early 1960ies. One can find the whole spectrum of political news and gossips that were actual and important those days. The collection is the smallest one, that is why a self-organizing map of size $10 \times 15$ is chosen. It is not easy to evaluate the results of LabelSOM method, but according to research and experiments described in [20] and [21] we consider it to be the method the results of which one can rely upon. Let us consider the map (see Figure 5.1) that we obtained.

5.1.1 Information provided for a unit

Let us consider, what kind of information we can get for a unit. In the previous chapters we described four layers of lexical items: LabelSOM labels, keywords, GATE annotations, context words of GATE annotations, and as a separate characteristics – the best context of a GATE annotation. Table 5.1 provides the whole spectrum for a unit
We get “space”, “woman”, “women”, “insurance”, “scientists” from the Label-SOM method and rather synonymical to them “valya”, “valentina” (the first woman in space), “vostok”, “cosmonaut”, “bykovsky” as keywords. GATE gives us the following persons: “valentina vladimirovna”, “scientists”, “vi”; locations: “tibet”, “moscow”, “russia”; organizations: “women astronauts”, “space”, “nasa” and the date “1960”. For each GATE annotation one can find its best context and best context words in the table.

It is possible to tell, what the unit is about without reading documents. It is even possible to precise some aspects. So, the unit is about the first woman in space (Label-SOM), more exactly, her name is Valentina (or Valya) and she is a cosmonaut of Vostok (Keywords). GATE tells us, that the full name is Valentina Vladimirovna, that the full “name” of Vostok is Vostok VI (although it is more or less a GATE mistake to classify “vi” as a person, but in our case, it is useful). The best contexts of GATE locations show, that Moscow and Russia are directly connected to the Tereshkova’s flight and that Russia’s space technologies are good developed. A GATE annotation (category Organizations) “nasa” let us get to know, that NASA is rather sceptic as to women in space. Due to the GATE annotations “scientists” and “tibet” we see as well, that the whole world still doubts the great possibilities of space travels and technologies.

The conclusion is, that using these four labelling layers, characterizing a unit, it is possible to get enough information about documents. We see that GATE annotations provide details (although we select three most important annotations of each category). In our example, almost all GATE annotations are able to characterize the whole unit, but very often, they give details that are not interesting. For example, “tibet” is rather unimportant for the unit. The best context of an annotation is always useful, if one needs to get details about some notion. The best context characterizes a word always in a correct way. The best context words, at the same time, are often difficult to refer
5.1.2 Brief description of the map

When we examine the GATE annotations for units, we notice some annotations of the category “Date”. Almost all the dates are obviously in the interval between 1959 and 1963.

The left upper corner includes documents about the Soviet Union, such LabelSOM labels as “krushchov”\(^1\), “pravda”, “nikita”, “izvestia”, “moscow”, “kremlin” assure us that this block of news is about the Crushchov’s politics and the situation in the country: “speculation”, “production”, “tons”. This rather big Soviet Union cluster, to

---

\(^1\)All words of the Time Magazine collection are written in capital letters. That is why the proper names in the results of our experiments do not start with a capital letter.
<table>
<thead>
<tr>
<th>LabelSOM</th>
<th>Keywords</th>
<th>GATE annotations and contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Persons</strong></td>
<td><strong>Locations</strong></td>
<td><strong>Organizations</strong></td>
</tr>
<tr>
<td><strong>Best context</strong></td>
<td><strong>Best words</strong></td>
<td><strong>Best context</strong></td>
</tr>
<tr>
<td>valentina vladimirovna</td>
<td>tibet</td>
<td></td>
</tr>
<tr>
<td>all the world’s women were in chains before valentina vladimirovna tereshkova</td>
<td>world's, khrushchov, dear</td>
<td>many scientists doubt astronaut gordon cooper’s report of seeing trucks on the road and smoke coming out of chimneys in tibet</td>
</tr>
<tr>
<td>many scientists doubt astronaut gordon cooper’s report of seeing trucks on the road and smoke coming out of chimneys in tibet</td>
<td>call, lost, ground</td>
<td>scanned the big story from moscow</td>
</tr>
<tr>
<td>vostok vi</td>
<td>russia</td>
<td></td>
</tr>
<tr>
<td>vostok vi</td>
<td>controls</td>
<td>the duo flight once again proved that russia is at least two years ahead of the u.s.</td>
</tr>
</tbody>
</table>

Table 5.1: Lexical items provided for a unit.
which one can refer a number of units \((0/0)^2 - (2/3) - (3/0)\) consists of some subclusters. One of them is realations of the USSR with its neighbor China \((3/0)\). Several units near this one, namely \((3/1), (4/0), (4/1)\) are about China. A label “atomic” let us understand that a question of atom weapon of China is considered.

Using the keywords method we get such words as “chemical”, “enemy”, “production”, “embassies”, “victims”, “commissars”. Using GATE we get a number of important locations such as “russia”, “moscow”, “siberia”, “china”, “peking”, “spain”, “britain”, “burma”. The persons selected by GATE provide a good characteristics for the cluster: “nikita khrushchov”, “kozlov”, “agents”. The organizations defined by GATE help to understand the content as well: “central committee”, “communist party”, “tass”, “soviet union” (“soviet union” is always put into a category “Organization” and not into “Location” because of a word “union”). As to the context, we get a good charcteristic for “khrushchov” – “enthusiastic”, “plug”, “equally” and a very correct feature of the Soviet Union of those days – “intellectuals”, “evgeny”, “shyest”. Berlin is associated with “frustration”, and China with “communists”. One of the neighbor units about international relations tell about Spain in terms of “coast”, “hopes”, “Atlantic” and about US in terms of “embargo”, “decides”, “bear”. The lower left corner, consisting of units , reflects the situation in Vietnam. One can notice two different topics here. The units about Vietnam are labels with semantically different words, for example, “combat” and “helicopter” versa “buddhists” and “monks”. The reason is that the news is about the religious conflicts and war. The cluster on Vietnam consists of two subclusters, namely an internal religious conflict, as well as documents on the Vietnam war. The keywords provide the same conclusion. The words “pope”, “vatican”, “holy” give a user an idea about a religious topic, and “violations”, “tank”, “hostilities” tell about a war conflict in the region. For this cluster GATE provides locations such as “viet nam”, “viet cong”,

\footnote{with \((x/y)\) we denote the unit in column \(x\), row \(y\)}
one finds “asia”, “u.s” and “soviet” several times (the neighbor units are labels with “tibet”, “saigon”). The context gives several interesting details, for “u.s.” – “influence”, “expressed”, “policies’; “asia” - “army”, “communist”, “party”; “viet nam” – “agreed”, “elections”, “finally”, “offer”, “malaysia”, “asked”. Between Vietnam problems and the Soviet Union – China relations, one can find a cluster about British and French nuclear forces. Such labels as “gaulle”, “nuclear”, “polaris”, “britain” assure us in that. Keywords method gives additional information with the words “strategic”, “planning”, “naval”, “submarine”. Locations found in the cluster are “france”, “britain”, “norway”, “denmark”, “western europe” that reflect the multinational idea of the nuclear plan. As to the context of separate annotations one can notice that France is described with “support” and “negotiations”, Geneva with “conference”, “europe”, “world”, “organization”.

The lower right corner tells about complicated situations in a number of countries in SE-Asia: Laos (2/12), Hong Kong (3/11), Cambodia and Thailand (4/12), and about a crisis in Congo and involvement of several powers in it (7/12), (7/13), (7/14). The keywords add the type crisis picture providing such words as “kill”, “revolt”, “worried” and “chaotic”. Using GATE one finds all the countries that were already marked by LabelSOM, and considering context of annotations, one finds a close to life connection “france” – “russia”, “congo”, “operation”.

The upper right corner is devoted to “the bloody Himalayan border dispute between China and India”. Between China/India problem and the news from the Soviet Union, one finds to some extent political scandal news, to some extent gossips about a Soviet spy, who is a first secretary of the Soviet Union. This cluster is described in terms of keywords as “enemy”, “truth” (it is mentioned that it is difficult to find the truth), “experts” (some scientists, experts seem to be spies). After a deeper analysis with GATE
annotations one finds personalities that could be spies (“Oleg Penkovsky”, “Giuseppe Martelli”), and the context provides a lot of information of a situation with Penkovsky - “trial”, “russia”, “slender”.

If we give a look to a central region of the map, we notice some significant clusters, for example, units (7/3) – (8/5) and (9/4) are about McMillan program and elections in England. The labels “workers”, “tories”, “market” characterize the documents that were put into these units very well. The keywords support this idea: “britons”, “debate”, “voters”. Among persons of this cluster there are “winston churchill”, “harold macmillan”, among locations one finds “london”, “britain”, “liverpool”, “utopia”. The context of an annotation “harold macmillan” is “elections” and “encourage” in one unit and “big”, “lost” in another.

Near political situation in England there is the political situation in Germany (units (4/2) – (6/3)), that is deduced from keywords “germany”, “adenerau”, “chancellor”, “bonn”, “wall”. So, e.g., one can read about a new chancellor of West Germany and compare his/her opinion to one of Konrad Adenauer. Keywords for this cluster are “votes”, “prestige”, “pact”, “representatives”. GATE annotations found are “conrad adenerau”, “wladyslaw gomulka”, “walter ulbricht” and “west germany”. From the context analysis we know that Wladyslaw Gomulka is connected somehow with “political” and “regime” and Adenauer is “chosen”, “champion”, “successor”.

5.1.3 LabelSOM labels versa keywords

Let us try to understand what is the difference in the results between the labels of the LabelSOM method and of the keywords. Let us examine the following example of two neighbor units (namely, units (7/0), (8/0), which are a subcluster of the multilateral nuclear program cluster). One of these units (see Figure 5.2) includes one document
and is characterized by the LabelSOM labels “britain”, “nato”, “fleet”, “ships”, “develop”. Another unit includes seven documents that are described by “nato”, “force”, “nuclear”, “solaris”, “allies”. The Keywords method provides the following words for these units: “marshal”, “demand”, “quietly”, “naval”, “pledge” and “strategic”, “sceptical”, “planning”, “submarine”, “massive” (see Figure 5.2). One can say, that the LabelSOM labels and keywords are rather synonymical and both methods are relevant for at least considered two units. To know more about the difference of methods, we read documents that were put into the units and find an interesting thing.

As already mentioned, the units discuss the French British nuclear program. The words of the LabelSOM provide objective information. The keywords, on the contrary, find information that is not of the priority importance, but that is far from being not important. A main topic of a document T308.html is possible problems of the program: “Can a ship fight effectively if one-third of its crew is Portuguese, one-third Belgian, and one-third, say, Danish?” These words belong to a marshal Montgomery. This fact is very good reflected with the keywords “marshal” and “sceptical”.

5.1.4 Comparison of two units

It is difficult to find borders of a cluster. Sometimes it is impossible to categorize a single unit as a part of this or that cluster. It happens as well, that two clusters are extremely good separated from each other, but are situated near by on the map. In this case, a question is, why these clusters are neighbors. One can solve this problem with the help of the comparison of both clusters or units.

Units are the smallest parts of the Self-Organizing Maps. They could be organized into subclusters, and subclusters in their turn, create clusters. We make several experiments comparing areas of a map, but we start with the smallest particles, two units.
Figure 5.2: LabelSOM labels and keywords.

The selected units are (0/2) and (0/3) on the map. The documents that are under consideration: for the first unit – T093.html, T311.html, T171.html, T245.html and for the second – T262.html, T399.html, T382.html, T134.html. While comparing two units, we review all the documents that are included in order to give a better explanation of the labels. When we compare more complicated structures, we will not examine all original documents in details and will only analyse the results provided by the labelling methods we use.

The first unit includes four documents on the following specific topics: an unsuccessful attempt of a Soviet young man to flee illegally from the Soviet Union; a spy in Moscow; Khrushchov is admired by American jeans, potato chips and corn flakes; Cuba breaks all contacts with the world for 24 hours. The second unit, that consists of four documents is about spy games and science: spy duet of Punkovsky and Wynne; an Italian nuclear scientist Giuseppe Martelli is a spy; spies in West Germany; Soviet
biologists have made an interesting discovery in Siberia.

Labels of LabelSOM method that are put into the first unit are “izvestia”, “soviet”, “comrades”, “moscow”, “pravda”. The second unit is characterized by “moscow”, “soviet”, “london”, “agents”, “intelligence” (see Figure 5.3). The words “moscow” and “soviet” are common for both units. We can examine not only the first five words, but, say, 15 words of the units (“izvestia”, “soviet”, “moscow”, “comrades”, “pravda”, “khrushchev”, “trip”, “celebration”, “giant”, “greet”, “shut”, “show”, “nikita”, “russia”, “missile” versa “moscow”, “soviet”, “agents”, “london”, “intelligence”, “espionage”, “russian”, “british”, “russia”, “izvestia”, “radio”, “scientists”, “tale”, “machinery”, “story”). We see that there are no more common words using the LabelSOM method in this case.

The keywords are “stamps”, “jet”, “telling”, “ugly”, “style” and “truth”, “happy”, “film”, “taught”, “mrs”. We have not a single common word comparing keywords of two units. The word “stamps” is frequently occurring in a document T093.html. Interest to stamps was the main reason, why an Ukrainian boy tried to cross the Soviet border illegally. The word “truth” characterizes the document T262.html. The spy scandal, where it is not easy to find the “truth”, and each accused tells his/her variant of truth.

The GATE annotations for the units give more detailed information about events (See Figure 5.4). For the first unit we get:

- “Person” : khrushchev, young russians, u.s. agents;
- “Location” : moscow, outside, havana;
- “Organization” : soviet press, maidison avenue, technical school;
- “Date” : November, 1959, 02/15/63.

For the second unit:
Figure 5.4: Time Magazine collection. Comparison of two units. Semantic zooming.

- “Person”: penkovsky, wynne, packages;
- “Location”: moscow, london, lido;
- “Organization”: courtroom, soviet scientists;
- “Date”: july, november 1960.

Having some background knowledge and GATE annotations, one can deduce what the documents are about. Using context analysis, it is possible to get several details, e.g. the characteristics of khrushchov “enthusiastic”, “plug”, “equally” is rather exact. It is not difficult to imagine what happens if one gets the following details for the persons: “oleg penkovsky” - “trial”, “russia”, “slender” and “martelli” - “declared”, “charges”, “guilty”. As it already was supposed, the keywords provide more specific information in comparison to the LabelSOM labels.
5.1.5 Comparison of two clusters

We choose a left upper corner of the trained self-organizing map on Time Magazine collection. This part of the map is about the Soviet Union, its inner affairs and international relations. The problem of the cluster definition is not easy. Its borders are not obvious. In this case we can say that we have one big cluster, that consists of several smaller subclusters. Analyzing the labels we obtain from the LabelSOM method, we decide that six units ((0/0) – (1/2)) belong more or less to a subtopic “Inner policy of the Soviet Union”, and below situated units ((2/0) – (3/2)) are connected with the international affairs of the country. Such a decision is based on the labels “committee”, “pravda” (one of the newspapers of the Soviet Union), “central”, “production”, “russia”, “tons”, “izvestia” (media channel), “kremlin” for the first cluster. And “chinese”, “treaty”, “turkey”, “peking”, “geneva” for the second cluster (see Figure 5.5. In order to distinguish the clusters, one of them is marked with the green colour, another one is yellow).

If we compare the LabelSOM labels of two clusters (see Figure 5.5) we get the common words: “china”, “khrushchov”, “kremlin”, “moscow”, “soviet” and a larger set of different words: “agreement”, “atomic”, “bomb”, “capitalists”, “chinese”, “comrades”, “conference”, “congress”, “delegates”, “editor”, “evgeny”, “geneva”, “intellectuals”, “izvestia”, “lenin”, “missiles”, “moon”, “nato”, etc.

If we compare these two clusters according to the keywords we get two sets of words. The common words for both clusters are “enemy” and “warning”. These two keywords give a good description for similarity of clusters about the situation of mistrust and tensions. The different terms include “addmitted”, “bank”, “check”, “chemical”, “chou”, “commissars”, “experts”, “hopes”, etc.

The next step is to compare clusters according to the GATE annotations that could

55
Figure 5.5: Time Magazine collection. Comparison of two clusters.
Figure 5.6: Time Magazine collection. Comparison of two clusters. LabelSOM method.
be found in the clusters. The method has a strong advantage: it is possible to get all the words that refer to the same semantic category, for example, find all persons or locations in a cluster. These categories are especially important for Time Magazine collection, full of political leaders and places. We see (Figure 5.7) that the common “Persons” is Nikita; “Locations” - Italy, Russia, Peking, Berlin; “Organizations” - WHO, Soviet Union; “Date” - May, 1959, March 22.

Figure 5.9 shows a summary. It includes all the information that we got about the two areas (the only thing, that we do not take into consideration, is separate context words of GATE annotations): LabelSOM labels, keywords, GATE annotations, best context of GATE annotations. LabelSOM labels and keywords are divided into three parts according to their importance, i.e. according to their weights. Words with the biggest weights are highlighted with red and the less important words are green. Word, that occur between these categories, are marked with the blue color.

5.1.6 Comparison of a unit and a cluster

Let us compare a cluster and a unit. An interesting task is to compare areas that are thematically similar, but have some significant difference. We take the cluster that we have already compared to another cluster, units (0/0) – (1/2) and a unit (3/3). The major topic is the Soviet Union, the cluster we have already described above in details. The unit we compare with the cluster concerns the Soviet Union, but it is the only unit on the map that tells directly about “khrushchov’s thaw”. The main topic of the unit is a young Soviet poet Evgeny Evtushenko. The LabelSOM labels are “poet”, “writers”, “intellectuals”, “khrushchov”, “evgeny”. The only common label of the unit and of the cluster is “khrushchov”. The Kaksi keywords of the single unit are “crackdown”, “evgeny”, “addmitted”, “houses”, “commissars”. A word “crackdown” is an undirect
Figure 5.7: Time Magazine collection. Comparison of clusters using GATE annotations.
characteristics of “intellectuals”. The document \textit{T218.html} tells about “the crackdown onevtushenko and other maverick intellectuals”. Common persons which we find in the unit and in the cluster thanks to GATE are “nikita chruschov”, common location are “russia”, common organization “soviet union”. On the context level we find the biggest difference between the unit and the cluster. One of the most suitable context characteristics of the unit is “soviet union” - “intellectuals”, “evgeny”, “shyest”. At the same time the most characteristic feature of the cluster is “soviet union” - “leadership”, “palace”, “future”.

5.1.7 Investigation of a map area

Let us select the already described area of the left upper corner. One of the most important categories is “Person”, and if we analyse an area according to persons (see Figure 5.8) we get “khrushchev”, “nikita”, “nikita khrushchev”, etc.

One of the main characters of the area is Nikita Khrushchev. If we investigate, in which context he could be found, we get 15 text fragments. The contexts are weighted according to words, that occur in them. Although each context weight is normalized to a number of words of a context, one notices, that the longer a context is, the more is its context weight.

It is useful not only to get a full context of a word, but to decide, which one is the best.
Figure 5.8: Time Magazine collection. Investigation of a map area. Weighted context.
According to the LabelSOM method, both areas deal with Moscow, Khrushchev, Soviet, China. Area 1 focuses on (LabelSOM method) nikita, editor, izvestia, moon, production, blast, rusia, steam, congress, tons, kremlin, comrades, pravda, capitalists; and area 2 on geneva, peking, poet, text, chinese, writers, turkey, treaty, agreement, atomic, intellectuals, bomb, nafo, evgeny, plane, delegates, momonara. Area 1 focuses on (Keywords method) CHEEK, POOL, TERKIN, LUNIK, FERTILIZER, SPITE, TVARDOCKY, STAMPS, VASILY, CHEMICAL, UPS, SNAPPY, JET, CHEMISTRY, TELLING, TOPICS, UGLY, PIONEERS, STYLE, VIRGIN, CAPITALISTS, FUNSTON, SCULPTOR, PAINTERS, SUITS, and area 2 on GROMYKO, WRITERS, PUBLISHING, CHOU, INSPECTIONS, WOLVES, YANG, SUBS, NOVY, BURY, TSARAPIN, VERMERSCH, MICHAEL, UNTRUTH, TURKEY, PALAIS, PANTS, WRIGHT, PACT, USHERED, JUMP, TENG, RUMANIAN, PROBE, CASTRO, SATELLITE, LENIN, SUSLOV, KIEN, YENAN. Both areas tell about the same following locations: BERLIN, CHINA, MOSCOW (MOSCOW ALSO PURGED THE FORMER PREMIER OF THE TERRITORY FROM THE LOCAL PARTY'S CENTRAL COMMITTEE), PEKING (PEKING'S DELEGATES TO THE COMMUNIST WORLD PEACE COUNCIL IN WARSAW NOISILY CONDEMNED RUSSIA'S "PEACEFUL COEXISTENCE" POLICY), RUSSIA (PEKING'S DELEGATES TO THE COMMUNIST WORLD PEACE COUNCIL IN WARSAW NOISILY CONDEMNED RUSSIA'S "PEACEFUL COEXISTENCE" POLICY). Both areas are about the same personalities: KHRUSHCHEV (KHRUSHCHEV TAKES NONE OF THE BLAME FOR THE FIASCO).

In these two areas one can get information about the following organizations: SOVIET PRESS, Persons and events are anchored to the time period of 1959, MARCH 22, MAY,.

Area 1 includes the following GATE annotations:
"Date": 01/11/53, 02/15/53, 1922, 1930, 1963, 1964, 1970, DECEMBER, MAY 28, NOVEMBER,
"Location": HAVANA, KAZAKHSTAN, NEW YORK, NOT, OUTSIDE, SIBERIA, SPLIT, UKRAINE,
"Organization": CENTRAL COMMITTEE, GENERAL FOODS, LIKE, MADISON AVENUE, ТASS, TECHNICAL SCHOOL, UPS, US COMMUNISTS, VIRGIN, VOLATILE PRESS, WALL, WARSAW, NOISILY, WHO,
"Person": CARRY OUT, DON, FUNSTON, NIKITA, TUNG POLITELY, U.S. AGENTS, U.S. USES, U.S. WEST, WEST, WILL ATTEND, WILL SPEND, YOUNG RUSSIANS,

And area 2:
"Organization": CHINESE, EAST, INDO-CHINA, LA PASCIONARIA, LADIES, MOSCOW WRITERS UNION, NOT, ON, PACT, RUSSIA AS, SOCIAL COUNCIL, THERE, TRAIN, TREATY, WEST GERMANY, EARLY, WEST GERMANY, STILL, WEST GERMANY

Figure 5.9: Time Magazine collection. Comparative summary of two clusters.
5.2 Collection Help Government (help.gv.at)

5.2.1 Brief description of the map

The collection Help Government is quite different from the Time Magazine collection. First of all, it is bigger and includes 1026 documents. We train a self-organizing map with $15 \times 25$ units (see Figure 5.10).

The documents are mostly in German, but one finds documents or paragraphs of documents in English, French, Turkish, languages of former Yugoslavia. The reason is, that the web site help.gv.at was developed to help people to find answers to questions of social and economical character. The significant part of such people are foreigners, who could have problems with the German language.

The left upper corner (units $(0/0) – (1/1), (1/2)$) of the map is dedicated to the ques-
tion how to open an enterprise. Such labels as “unternehmungsgründung”, “franchising”, “arbeitnehmerinnen”, “beschäftigungssformen” prove this idea. Keywords “GmbH”, “betriebsanlagengenehmigung”, “betriebsanlagen” characterize the cluster additionally. Among locations there are “Gründerinnenzentrum”, “Austrian”, “Steiermark”. An annotation category “profession” provides words “Unternehmer”, “Angestellte”.

An important remark must not be ignored. One of the Jape rules, that use GATE to annotate lexical items, is, that if a single capital letter is followed by a point, then this letter and a following word create a structure that is an abbreviation of a first name and a following it last name. So, that is why this structure is considered by GATE to be a “person”. For example, “B. Befreiungen” and “B. Barzuschüsse” are categorized as persons. Another interesting fact is that there is a number of words that are considered by GATE to be first names. In such a case GATE takes a first name and a word that follows it to annotate them as “person”. That is why “Thema Firmenbruch” and “Thema Aussenhandel” are persons.

The topic of work takes another turn a little bit lower on the map and a cluster is about relations between an employer and an employee, such words as “arbeitsverhältnis”, “arbeitnehmer”, “arbeitgeber” are very typical for a number of units beginning under the enterprise cluster and till almost the bottom of the map. This cluster is one of the biggest on the map and has several subclusters. It is even difficult to define its relative borders, that is why we describe it according to its subclusters. One of subclusters includes units (2/0) – (3/2) and is generally characterized by “entgelt”, “allgemeines”, “anmeldung”. This subcluster is about the basic things concerning work: types of work contracts, fees. A typical document of this area, for example, is Seite.090000.html “Allgemeines zu Rechten und Pflichten der ArbeitgeberInnen und ArbeitnehmerInnen” (“General positions as to rights and duties of an employer and an employee”). The
unit (2/3) is about migrating workers and is labeled with “saisoniers”. The units (4/0) and (4/1) tell about the specific situations connected with work and are labeled with “dienstreise” and “krankmeldung”. The big subcluster a little bit lower (units (6/2), (6/3), (7/1), (7/2), (7/3), (8/2)) is labeled by words “karenz”, “anspruch”, “dauer”, “kinder”, “kindes”, and it is not difficult to tell that the main topic is employees with new-born or small children. The keywords for this large cluster are different according to a subcluster they refer to. E.g. in the upper regions of the cluster one finds “Dienst”, “unbefristetes”, “Entgelt” (some keywords are common with the LabelSOM labels), a little bit lower words “Urlaub” and “Kranke” and their derivates – “Urlaubsjahr”, “Urlaubsantritt” and “Krankenversicherung”, “Krankenschein” – dominate. The GATE annotations for the area are “Gebietskrankenkasse”, “Krankheitskosten”. A cluster with the GATE annotation “Pflegeheim” is nearby and is characterized by a context words “Kinderbetreuung” and “Kosten”.

The subcluster concerning children is characterized with “Mutterschutzgesetzes”. Persons of this area are often “Mutter” and the profession mentioned is “Arbeitnehmerin”. In the category “Time” one finds “Elternzeit”, “Teilzeit”. “Karenzurlaubgesetz”, “Elternkarenz”, “Mutterschaft” refer to the topic perfectly, but are defined by GATE as “persons”.

The lower left corner is devoted to possibilities to get help, answers to specific questions on-line or to download forms. The units (0/13) – (2/10) are characterized by “help”, “information”, “internet”, “service”. One notices that a datum 2004 or 22.12.2004 occurs very often, and not only in this part of the map. This date shows how old is the information provided by a web site. The keywords describe the area with “Server”, “Fragen”, “Online”, “Internet”. It is the region of the map, for which it is not informative enough to use GATE annotations. Using GATE we have got a
number of widespread notions like “Heirat”, “Zivildienst”, “Wehrdienst”, etc. In order to understand what are they for, we examine their context and find out, that a lot of notions have no context at all, others possess the not informative context. This happens, because the documents which are put onto this part of the map, are more or less textless links to questionnaires or links to another web-sites.

One the bottom of the map there are two clusters about education and protection of customers rights. The typical LabelSOM labels are “ausbildung”, “lebensjahr”, “förderungen” and “konsumentenschutz”, “sozialversicherung”, “sicherheit” corresponding. The keywords for the two clusters are “schulische”, “bestehenden”, “StudentInnen” and “Vertretung”, “Versicherung”, “Versteuerung”. Using the context analysis we get “Erasmus-Programmes” – “StudentInnen”, “Rahmen”, “absolvieren”.

The upper right corner looks very undefinable. We find words that are not able to characterize a cluster within the map, the words are too typical for the whole text collection, e.g., “behörde”, “gesetzlichen”, “antrag”, “kosten”. Let us examine several documents that occur in this cluster (units (21/0), (22/0), (23/0), (24/0), (22/1), (24/1), (22/2), (24/2), (20/3)). Signature cards are discussed in \textit{Seite.2800100.html}, Genehmigungsverfahren in \textit{Seite.520200.html}, Gebrauchsmuster (ein kleines Patent) in \textit{Seite.713000.html}, Individuelles Bakkalaureats-, Magister- oder Diplomstudium in \textit{Seite.160700.html}, \textit{Seite.1130004.html} is an index of terms beginnig with the letter “D” (We find there explanations of notions “Darnachhaltung”, “Delegierung”, “Delikt”, etc.) None of these words, that are keywords, is noticed by the LabelSOM. It happens, because all of them are rare words in the real life and are described in one, maximum two documents. The Internet topic dominates within kewords for this cluster: “link”, “Web”, “downloaden”. The reason is that several documents of this cluster describe how it is possible to communicate with authorities using the Internet and new tech-
nologies. GATE notices a number of topics discussed in the documents of this cluster: "Adoption", "Bankomat" (defined as "persons"), Organizations found are "Magistrat", "Arbeitsgemeinschaft", "Arbeitsinspektorat". While examining the context of separate notions we find "Software" connected with "wesentlich" and "Schutz", "EU" is associated with "Euro", "Heirat", "Arbeitnehmerveranlagung".

A little lower situated cluster (units (22/4), (23/4), (24/4), (22/5), (23/3), (24/5), (22/6), (23/6), (24/6)) has the same trend: it is not easy to find out, what it is about. The labels are "dokumente", "formular", "gebühren", "nachweis", "lichtbildausweis", "mitzubringende". After reviewing documents of this cluster, we make a conclusion that the cluster is about obtaining documents, for example, driving license (Seite.040500.html), bank card (Seite.622000.html), etc. The documents of the cluster tell, what authorities need, whether one has to pay a fee and fill out a form. These instructions, namely, what to do in order to get this or that, are described by the LabelSOM labels. The keywords provide more precise information to the topic of obtaining documents and giving information what kind of documents exist: "internationaler", "akademischer", "alternativer". The words "formulare" and "online" are present here too. The largest category determined by GATE is "LOC" (locations): "Krankenhäuser", "Stadtgemeinde", "Fachhochschulen", "Pflegeheim", "Pfarrikirche", etc. The more detailed context analysis gives us such specific information like "Standesamt" – "Dokumente", "mitzubringende", "Heiratsurkunde"; "Staatsangehörige" – "Hinweis", "österreichischen", "EU".

In the central area of the map there are a couple of big clusters. One of them is about children-parents problems, for example different kinds of fees, special situations of not paying fees, answers to a question, when parents get paid. This cluster includes the following units: (8/19) – (11/24) the labels demonstrate it very good: "kind", "eltern", 

67
“soziale”, “sicherheit”, “geburtsurkunde”, “kosten”, “hilfe”. The keywords precise the description, given by the LabelSOM labels, with the following words: “behinderte”, “finanzielle”, “Schule”, “Vermittlung”. GATE provides that several of these units have “Eltern” and “Geschwister” in the “PROF” (profession) category, among organizations there are “Ehegemeinschat”, “Hausrat”, “Familienreferat”, “Familienverband”. In the category “TIME” one finds “Schulzeit”.

Another not so large centrally placed cluster (units (12/8) – (14/9) and (14/7)) includes a term “ausland” in its every unit. The cluster is about everything that is connected to the crossing the national border, which representatives of Austria could be found abroad and what foreign officials are in Austria. The keywords are “Aufe
enthalts Genehmigung”, “magistratischen”. How we already decided, documents of this cluster help foreigners in Austria or Austrian people abroad. That is why this cluster is the reachest with the GATE locations: “Australien”, “Deutschland”, “Kanada”, “Norwegen”, “USA”, etc. As we see, almost all the countries are not EU countries and are special cases for foreigners. As to the context, we find that “Wien” is con
nected with “AuslandsösterreicherInnen”, “Abmeldung”, “Aufenthalts Genehmigung” and “Österreich” with “Anmeldung”, “Standesamt”, “mitzubringen”.

5.2.2 Comparison of two units

In the Time Magazine collection we compared two units that were neighbors but had not very much in common. Let us now take two units from the central area of the map trained on HelpGov collection (2/7) and (2/8) that refer to the same cluster (about working people with children) and make an assumption that the units are very similar.

The first unit includes the following documents: Seite.400000.html, Seite.970004.html and Seite.870700.html with the titles “Karenz”, “Entgeltfortzahlung
bei Krankheit”, and “Urlaubsanspruch bei Karenz”. The second unit includes the documents Seite.409999.html, Seite.171600.html, Seite.400200.html, Seite.420700.html and Seite.990034.html with the corresponding names “Häufig gestellte Fragen und Antworten zu “Karenz” (FAQs)”, “Wiedereinstieg nach der Karenz”, “Teilzeitbeschäftigung neu”, “Urlaubsanspruch bei Karenz” and “Karenz”. It looks like that both units are about just the same, but the SOM algorithm have not put all the documents into the same unit. After comparison, we will try to answer the question: what is the difference between the units?

The LabelSOM labels for the first unit are “karenz”, “wochen”, “arbeitnehmerin”, “anspruch”, “dauer”; for the second unit the LabelSOM provides “karenz”, “arbeitgeberin”, “kindes”, “wochen”, “anspruch” (see Figure 5.11) and we see that there are some common labels. The keywords (see Figure 5.11) do not possess common words but support the idea, that the units are about “karenz”: we find “Mutterschutzgesetzes”, “Arbeitnehmerin”, “Urlaubsanspruch”, etc. GATE gives a number of useful annotations, although they are sometimes put into a false category (see Figure 5.12). For the both units such annotations are common: “Karenzurlaubsgesetzes”, “Wirtschaftskammer”, “Arbeitsinspektorat”, “Wien”, “Elternteilzeit”, “Eltern”, “acht Wochen”, “vier Wochen”. The context analysis seems very similar for the units. The thematical difference is practically absent, but if we examine the GATE annotations, that are different for two units more carefully, we notice, that the second unit has much annotations in the “DATE” category. After the analysis of the dates, one comes to a conclusion, that these dates are connected with the fact, that an employer has already a child. At the same time the first unit contains the documents concerning rather employers who is going to have a child, but still do not.
Figure 5.11: HelpGov collection. Comparison of two units.

Figure 5.12: HelpGov collection. Comparison of two units. Semantic zooming.
5.2.3 Comparison of two clusters

In the Time Magazine collection we compared two subclusters of one cluster. Let us try to compare two rather different clusters, although they are neighbors and have something in common. Our first cluster consists of units (23/9) – (24/12) and the second cluster of (21/9) – (22/10), (23/10) (see Figure 5.13). Generally speaking, one can tell, that both clusters are about how to obtain this or that document, what is needed and how a procedure is carried out.

The LabelsSOM labels that are shared by both clusters are “antrag”, “beispielweise”, “dokumente”, “geburtsurkunde”, “heiratsurkunde”, “nachweis”, “österreichischen”. There are fewer keywords that are shared by the clusters: “fragen”, “urkundlicher”, “magistrat” (see Figure 5.14). Let us make a comparison on the level of GATE annotations. The GATE categories, especially “PER”, “LOC” and “PROF” give us much more better characteristics of the clusters than LabelsSOM labels and keywords, that provide too general notions. Let us consider the table 5.2.3, which shows all the GATE annotations of the two clusters. We see, that the first cluster is much more focused on children (“Ferienzeit”, “Studienzeit”, “Lebensjahr”), words with the roots “Eltern” and “Schule” occur very often. Thus, cluster 1 is about different aspects of parents – children relationship, different cases of adoption, insurance, and education. The second cluster contains less documents, and hence, less GATE annotations. The GATE categories “LOC” and “ORG” help us to determine that the cluster is about special cases of education: special schools (e.g., “Heilstättenschule”) and organizations (e.g., “Sehbehindertenverband”).
<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>DATE</th>
<th>TIME</th>
<th>Exception</th>
<th>LOC</th>
<th>ORG</th>
<th>PER</th>
<th>PROF</th>
</tr>
</thead>
</table>

Table 5.2: Comparison of two clusters based on GATE annotations.
<table>
<thead>
<tr>
<th>Label1SOM</th>
<th>Label1SOM</th>
<th>Label1SOM</th>
<th>Label1SOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>inhaltlich ausgeartet in die bestehenden österreichisch</td>
<td>nachweis erteilt mittels la dokumente ausgeartet</td>
<td>amtlicher lichblausweis nachweis österreichischen ausland</td>
<td>amtlicher lichblausweis ausland staatsbürgerschaftsnachweis geburtsaufende</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Label1SOM</td>
<td>Label1SOM</td>
<td>Label1SOM</td>
<td>Label1SOM</td>
</tr>
<tr>
<td>österreich voraussetzungen genüber grund nachweis</td>
<td>österreich höratsurunde nachweis antrag dokumente</td>
<td>höratsurunde antrag beispielsweise gebühren staatsbürgerschaftsnachweis meldezettel</td>
<td>höratsurunde staatsbürgerschaftsnachweis geburtsaufende meldezettel</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Label1SOM</td>
<td>Label1SOM</td>
<td>Label1SOM</td>
<td>Label1SOM</td>
</tr>
<tr>
<td>österreich staatsbürgerschaftsdeskumente</td>
<td>meldezettel anmeldung staatsbürgerschaftsnachweis höratsurunde geburtsaufende dokumente</td>
<td>meldezettel geburtsaufende staatsbürgerschaftsnachweis heiratsurunde bestätigung formular</td>
<td>meldezettel geburtsaufende staatsbürgerschaftsnachweis personen dokumente staatsbürgerschaftsnachweis</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>11</td>
<td>4</td>
</tr>
</tbody>
</table>

Figure 5.13: HelpGov collection. Comparison of clusters.
Figure 5.14: HelpGov collection. Comparison of clusters according to key-
words.

Figure 5.15: HelpGov collection. Comparison of clusters according to GATE annotations.

Figure 5.16: HelpGov collection. Comparison of clusters according to context of GATE annotations.

Figure 5.17: Time Magazine collection. Comparison of clusters according to context of GATE annotations.
5.2.4 Comparison of a unit and a cluster

Let us compare a unit (9/24) with a cluster (10/22) – (11/24). A unit can be considered to be a subcluster of the cluster. The common LabelSOM labels are “generatien”, “konsumentenschutz”, “soziale”, “vertretung”. Common GATE annotations are “DATE” - “3.12.2004”, “22.12.2004”.

The dates show, since when the acts are promulgated. A GATE annotation of the unit and its keyword “Familienbeihilfe” is the most important for this unit. As to the whole cluster, the subcluster of which is the unit, the GATE annotations “Arbeitsplatz” (LOC) and “Bundessozialamtes” (ORG) are typical. In this case LabelSOM method provides much more better characteristics for the area than other methods. The reason is that the cluster has a good determined topic (protection of consumers rights and social services), but every document devoted to this topic, is about some specific concept (e.g., “Lehrvertrag”). Thanks to the LabelSOM labels it is not difficult to define, what the cluster is about, but it is quite difficult to get details.

5.2.5 Investigation of a map area

Let us select the first option of the menu “Annotations of the selected area”. We see the annotations of the semantic category of the selected area. For HelpGov collection one of the most important semantic categories are “Professions” and “Locations”. So, among professions we find “Herausgeber”, “Abonnent”, “Bürger”, “Vermittler”, etc. As to locations of this area, let us consider already weighted annotations (see Figure 5.19).

The third option of the menu “Annotations with context” (Figure 5.20) shows the annotations of a selected semantic category and gives a possibility to choose an annotation. The information provided for each annotation includes the full instance context of a word in the selected area, the weight for each word of this context and part-of-speech
In this experiment we did not eliminate stop words and come to the conclusion, that if all the contexts are of the same length (more than five words), eliminating of stop words is not so important for finding the best context. At the same time one can use different weights within the grammatical categories that are considered to be stop words, such as pronouns, prepositions in order to compose a much more precise linguistic summary.

An annotation word is not considered as a context word. The smaller a value of a weight is, the more important a word is.

Let us analyse what we get. We see, that for a GATE annotation “Bürger” the most important context words in this area are “Umzug”, “Amt”, “informieren”, as the area is about formalities that are to be done in case of a removal.
5.3 BankSearch DataSet 7.0

5.3.1 Brief description of the map

The BankSearch DataSet collection is the largest one of our experimental collections, and we train the map $25 \times 35$. The complexity is too high, and that is why not all of the experiments, that we have carried out for the previous two collections, are performed. We get the LabelSOM labels and GATE annotations for the BankSearch DataSet. In the experiments with the BankSearch DataSet one notices, that different GATE categories are important for different map clusters, namely, “Money” is important for the lower right corner, where we find documents devoted to finance; “Persons” and “Locations” characterize the upper right corner – we get names of sportsmen and locations, where competitions take place; Organizations are important for the left upper corner connected to the scientific space organizations, as to the left lower corner, LabelSOM method provides much more better labels, than Keywords method and GATE annotations.

Let us start traditionally from the left upper corner. We find there a big cluster (units $(0/0)$ – $(3/9)$ and $(4/0)$). The cluster is described with the following La-
Figure 5.20: HelpGov collection. Finding the best context.
belSOM labels: “gallaxies”, “astro”, “astronomy”, “telescope”, “moon”, “plantes”, “nasa”. It is not difficult to understand, that the cluster is about the NASA space missions, hubble telescope and recent news in the field of astrophysics. GATE annotations provide a lot of additional information. The most interesting semantic categories for this cluster are “Address” and “Organization”. We find a number of addresses, such as “http://astrosurf.com/skylink/”, “http://time.space.free.fr/”, “http://lisa.jpl.nasa.gov/” etc; organizations found are “Infrared Space Observatory”, “Stratospheric Observatory”, etc.

The lower left corner is devoted to problems in programming languages, e.g. Visual Basic, HTML. The area of the cluster consists of the units (0/22) – (10/34). The cluster is described with “operator”, “template”, “integer”, “pointers”, “char”, “vb.net”, “asp.net”, “applet”, “sort”. If we examine some documents of this cluster in detail, we find out that the document F0799.txt is, for example, a web site of a Visual Basic 6.0 forum, E0508.txt is a piece of a tutorial on C++. Among GATE annotations of this cluster one finds “Microsoft”, “Sun Microsystems”, “MS Office”, “VB” (all categorized as “Organization”), “Java” (categorized as “Location”). There are a lot of dates and persons, because of forum sites documents.

The right lower corner is dedicated to finance. The LabelSOM method provides the following keywords for the cluster: “withdrawal”, “deposit”, “offshore”, “discount”, “tariff”, “marketplace”, “overdraft”, “cheque”, “transaction”. The cluster lies within the area defined by a triangular (12/34), (24/34) and (24/26). The categories of “Money” and “Location” dominate among the GATE annotations. A number of world financial centers like “New York”, “Hong Kong”, “Singapore”, etc. could be found. The organizations of the cluster provide a good description for the cluster as well: “Citibank International plc”, “Financial Times”, “Banking CodeSavings”, etc. The annotations
of the “Percent” category are numerous too.

The right upper corner (units (19/0) – (24/4)) is about football. It is not so easy to define it. The first, what one notices, it is a number of countries: “holland”, “ukraine”, “denmark”, etc. One finds as well words “stadium”, “match”, “uefa”. Practically the whole cluster is about the World Cup 2002. GATE annotations of the “Location” category include all the countries that participated in the World Cup. The “Person” is an interesting category for this cluster. We find the most prominent soccer players, such as “Zidane”, “Luis Chilavert”, “Paolo Maldini”, “Raul”.

Under this cluster is another big sport cluster that is characterized with “marathon”, “championship”, “tournament”, “olympic”, “bike”. As a subcluster of this cluster one can name the units (20/24) and (21/24). Both of them are labeled with “extreme” and “surfing”. Among GATE annotations one can notice a number of organizations, such as football clubs: “Chelsea Football Club”, “England Arsenal”.

5.3.2 Comparison of two units

In our previous experiments we always choose two neighbor units. Let us take two units ((5/2) and (1/1)), that are not neighbors, but belong to the same cluster. The first unit includes 3 documents, the second unit consists of 18 documents. The LabelSOM method provides the following labels respectively: “hubble”, “galaxies”, “telescope”, “galaxy”, “fluctuation” and “hubble”, “observations”, “hole”, “asteroids”, “resolution”. We notice, that the word “hubble” is common for both units.

The unit (5/2) in terms of GATE annotations is characterized with a number of countries that participate in research and with a number of “Person”-scientists. There are not a lot of organizations mentioned in this unit, but those, that are present, are important: “Human Spaceflight Hubble Service”, “Time Warner Company”, “Royal
Observatory” and some other observatories. The unit (1/1) is rich with persons and dates.

As to one of the most important for both units words, “hubble”, it is annotated as “Person”. Although these units are not neighbors, it is extremely difficult to find a subject difference.

5.3.3 Comparison of two clusters

Let us compare two clusters (units (22/9) – (24/10) and (20/10) – (21/11) respectively), both of which are devoted to sport. The common LabelSOM labels are “championship”, “coaching”, “tournament”, and “youth”.

The GATE categories of “Date” and “Location” are very characteristic for both clusters (other categories are almost absent in the first unit): one can deduce when and in which country a championship or another sport event takes place. It is possible to find out which countries participate in it as well.

The categories of “Person” and “Organization” are characteristic for the second cluster, so one finds “Nottingham Forest Programme Club”, “Newcastle United FC”, “Coastline DS Ltd”, “Chelsea Independent Supporters Association”, etc. After the comparison, we come to the conclusion, that the topic of football is only presented in the second cluster, as to other sports, for example, rugby, surfing, they are shared by both clusters.

5.3.4 Comparison of a unit and a cluster

Let us consider the right lower corner (devoted to the financial news), and compare a unit (21/32) (consisting of 49 documents) with the cluster (units (22/33), (23/33), (24/33), (22/34), (23/34), (24/34), (21/33), (21/34)). The common LabelSOM labels
are “bonus” and “garanteed”.

GATE provides much more common annotations, first of all, in the “Organization” category: “Financial Services Authority”, “Legal & General Marketing”, “Financial Services Authority”, “Payment Protection Intermediary Service”. One can say, that the “Organization” category is the most important one for the unit. As to the cluster, we find numerous annotations of the categories “Money” and “Percent”. It is necessary to notice, that an annotation “Account”, shared by the unit and the cluster, is categorized as “Person”.

5.4 Summary

In this chapter we described our experiments with three different text collections. For the text collections the comparisons unit – unit, unit – cluster and cluster – cluster were performed according to four layers of precision: LabelSOM layer, keywords layer, GATE annotations layer and context of GATE annotations layer (for the BankSearch DataSet collection the comparison was made only according to two parameters – LabelSOM labels and GATE annotations). The detailed investigations of map areas were carried out for two collections.

We made a conclusion that different semantic GATE categories are important for different map clusters. The context words of GATE annotations provide good characteristics not only for an annotation they refer to, but for a unit as well.
Chapter 6

Conclusions

In this work we reviewed the SOM algorithm, the task of text clustering and the notions that are connected with it: text indexing, weighting, vector normalization, and stoplists. The methods to label units of the SOM were examined and analyzed. In this work the LabelSOM method and the keyword selection method, developed by K. Lagus and S. Kaski, are used. A special role in this thesis is played by GATE, a tool to process text in natural language according to semantic categories.

Two tools were developed, one of them is focused on the analysis of a map area (“Semantic description of a map area”), another one makes a comparison of two map areas (“Comparison of map areas”).

To evaluate the tools, experiments on three text collections (Time Magazine collection, HelpGovernment, and BankSearch DataSet) of different size (420, 1,024, and 11,000 documents) and different languages (English, German) were carried out. Comparisons of unit – unit, unit – cluster and cluster – cluster according to the LabelSOM labels, keywords, GATE annotations, and context words of GATE annotations were performed. We investigated separate map areas as well.

The results have shown that the LabelSOM labels provide the best general charac-
teristics for units. keywords could be considered to be a one layer deeper analysis of a unit. Usually, the LabelSOM method and keywords provide synonyms. The number of common words among LabelSOM labels and keywords were expected to be higher than it occurred to be.

One layer deeper than keywords (and two layers deeper than LabelSOM labels) are GATE annotations, that provide every unit with sets of words, which are put into semantic categories. So, for the English language these categories are “Person”, “Location”, “Organization”, “Address”, “Age”, “Offence”, “Drug”, “Time”, “Date”, “Percent”. The GATE application for the German language has a different set of annotations: date (“DATE”), exception (“Exception”), location (“LOC”), number (“NMB”), organization (“ORG”), person (“PER”), time (“TIME”), profession (“PROF”), percent (“PER”).

The last layer of labels is the context layer. It is created using GATE annotations of a previous layer. The context words provide, in the most cases, a good characteristics for a GATE annotation. The weight for this method was introduced in this thesis. The method is based on the Keyword selection method and the notions of type and instance contexts, which were introduced in this thesis as well.

While using all four layers of unit analysis (LabelSOM labels, keywords, GATE annotations and context of GATE annotations), one gets enough information to tell what a unit is about without reading documents of this unit. Hence, a problem for the future work would be to create more semantic categories. Another problem is that semantic categories are not only language dependent, but more or less domain dependent. So, e.g. for the Time Magazine collection the categories of “Location” and “Person” are the most important and the “Time” category does not play a special role. At the same time for the HelpGovernment collection a category of “Organization” was
extremely important, and the “Person” category provided mostly incorrect results.

It is difficult to compare the LabelSOM and Keywords selection methods with the GATE annotations method. GATE annotations of a unit are selected without paying attention to the weight of a word; already selected GATE annotations are weighted using the Keywords method, but it is possible that a unit is characterized with a GATE annotation that is not a good descriptor according to the LabelSOM and the Keywords method. At the same time GATE gives a possibility to get all lexical items that are of the same semantic category. To find a method to weight GATE annotations that would correlate with the question, how important a GATE annotation is for this or that unit, is a problem for the future work.

We created a comparative summary of two selected areas of the map. The summary is based on templates and uses LabelSOM labels, keywords, GATE annotations their best contexts. The next step is to add grammatical categories and a thesaurus to the tool in order to develop a summary which is much more closer to a natural language.
Appendix A

Notations

\( C = \{c_1, ..., c_m\} \)  training set

\( t_k \)  lexical item of a document \( d_j \)

\( w_{jk} \)  weight of a lexical item \( t_k \) of a document \( d_j \)

\( \#(t_k, d_j) \)  number of times \( t_k \) occurs in \( d_j \)

\( \#(t_k) \)  number of documents in \( Tr \) in that \( t_k \) occurs at least once (document frequency)

\( Tr \)  Training set

\( \#(t_k, Tr) \)  term occurrence frequency

\( n \)  all the terms that occur at least once in \( Tr \)

\( m_i = [\mu_{i1}, \mu_{i2}, ..., \mu_{im}]^T \in \mathbb{R}^n \)  vector of output units

\( x = [\xi_1, \xi_2, ..., \xi_n]^T \in \mathbb{R}^n \)  input vector

\( c \)  a winning unit (a winner)

\( t \)  current training iteration

\( h_{ci} \)  neighboring function

\( \alpha(t) \)  learning-rate factor (monotonously decreasing function)

\( \sigma(t) \)  width of a kernel (monotonously decreasing function)

\( U \)  unit
Goodness of a descriptor

\( G \)

Goodness of a context word within the type context

\( G_{gen} \)

Goodness of a context word within the instance context

\( G_{spec} \)

Goodness of a context word, based on the values \( G_{gen} \) and \( G_{spec} \)

\( G_{context} \)

A value, that defines the relation of \( t_k \) and other words within a unit

\( F^{clust} \)

Relates the word \( t_k \) to the whole collection

\( F^{coll} \)

Number of times \( t_k \) occurs in unit \( \ell \)

\( f_\ell(t_k) \)

Relative frequency of \( t_k \)

\( F_\ell(t_k) \)

Neutral zone on a map

\( B \)

Radius of a map area

\( r_0 \)

Radius of a neutral zone

\( r_1 \)

Quantization error vector

\( q_{ik} \)

Number of labels

\( \lambda \)

Thresholds

\( \tau_1, \tau_2 \)

Locations of vectors of nodes \( c \) and \( i \) on the map grid

\( r_c, r_i \in \mathbb{R}^2 \)
Bibliography


