VOL 1 (2017) NO 4 - 2

e-ISSN : 2549-9904
ISSN : 2549-9610

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

X-Similarity Comparison by Using Wordnet

Shahreen Kasim[#], Nurul Aswa Omar[#], Nurul Suhaida Mohammad Akbar[#], Rohayanti Hassan[*],
Masrah Azrifah Azmi Murad[**]

[#]Department Web Technology, Faculty Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia
[*]Faculty of Computing, Universiti Teknologi Malaysia, Malaysia
[**]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia , Malaysia
E-mail: nurulaswa@uthm.edu.my, shahreen@uthm.edu.my, ai120067@siswa.uthm.edu.my

Abstract— **Semantic web is an addition of the previous one that represents information more significantly for humans and computers. It enables the description of contents and services in machine readable form. It also enables annotating, discovering, publishing, advertising and composing services to be programmed. Semantic web was developed based on Ontology which is measured as the backbone of the semantic web. Machine-readable is transformed to ma-chine-understandable in the current web. Moreover, Ontology provides a common vocabulary, a grammar for pub-lishing data and can provide a semantic description of data which can be used to conserve the Ontology and keep them ready for implication. There are many that used in feature based in semantic similarity. This research presents a single ontology of X-Similarity feature based method.**

Keywords— **X-Similarity, Wordnet.**

I. Introduction

Semantic similarity is a determination between words in many natural language tasks such as word sense disambiguation, document categorization or clustering, word spelling correction, automatic language translation, ontology learning or information retrieval. Semantic similarity also computes the likeness between words; understand as the degree of taxonomical proximity. For example, “monitor” and “CPU” are similar because both are part of comput-er. Word also can be related in non-taxonomical ways. The approach in the semantic similarity, the meaning of a target text is inferred by assessing how similar it is to another text. It is called the benchmark text whose meaning is known. According to some measure of semantic similarity, if the two texts are similar enough, the meaning of the target text deemed similar to the meaning of the benchmark [1].

Ontology captures a certain view of world, support intentional queries regarding the content of database and re-flects the relevance of data by providing a declarative description of semantic information independent of the data representation [2]. Ontology is a representation of knowledge. It is a set of concepts within a domain and the rela-tionship between these concepts. Ontology is used in semantic web, system engineering, software engineering, bio-medical informatics and many more.

II. Literature Review

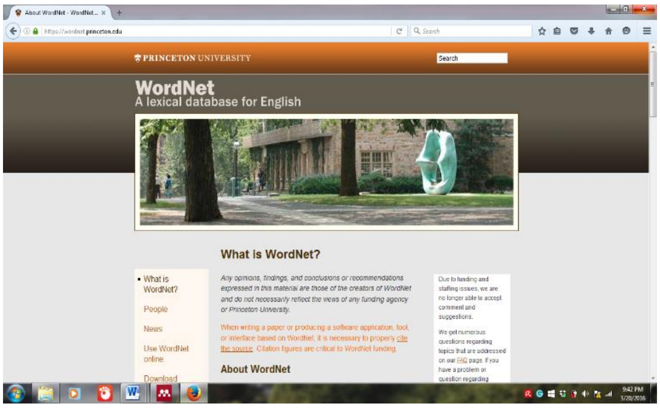
Ontology is a “An ontology is defined as a formal, explicit specification of a shared conceptualisation” which means that ontology is defined as a formal representation of concepts within a domain and the relationship be-tween those concepts [4]. Ontology is an effective way to share knowledge within controlled and structured vocabu-lary [5]. Many ontologies have been developed for various purposes and domains [6-8]. Furthermore, in reference to [9] ontology is built for some reasons such as sharing a common understanding of the structure of information among people or software agents, enabling the reuse of domain knowledge, making explicit domain assumptions, separating the domain knowledge from the operational knowledge and analysing domain knowledge. Besides that, ontology is also crucial in enabling interoperability across heterogeneous systems and semantic web applications [10].

Ontology is a type of knowledge-based that describes concepts through definitions that are sufficiently detailed to capture the semantics of a domain. A few ontologies such as the WordNet [8] have been used for semantic simi-larity. The WordNet is a lexical database for general English covering most generic English concepts and supports various purposes. Besides that, other ontologies are also used for the same purpose as the Unified Medical Lan-guage System (UMLS), that includes many biomedical ontologies and terminologies (e.g., MeSH, Snomed-CT) [11], and the International Classification Disease (ICD) family [6]. These ontologies are specifically created for the bio-medical domain that is different from WordNet.

There are several examples of general purpose ontologies available including: WordNet, SENSUS, and the Cyc knowledge base. The following section describes the general purpose ontologies as follows:

WordNet

WordNet is the lexical knowledge of a native speaker of English. The latest version of WordNet is v3.1 which was released in June 2011. WordNet has 117,659 synsets and 206,941 general concepts of different domains [12]. These databases are semantically structured in ontological ways. It also contains nouns, verbs, adjectives and ad-verbs that are linked to synonym sets (synset), where each synset consists of a list of synonym word forms and se-mantic pointers that describe the relationships between the current synset and other synsets (Hliaoutakis et al., 2006). Different types of relationships can be derived between the synsets or concepts (related to other synsets higher or lower in the hierarchy).



WordNet
A lexical database for English

What is WordNet?

- What is WordNet?
- People
- News
- Use WordNet online
- Download

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the creators of WordNet and do not necessarily reflect the views of any funding agency or Princeton University.

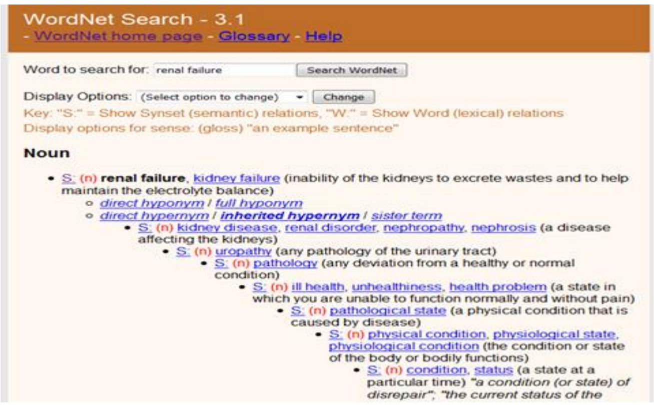
When writing a paper or producing a software application, tool, or interface based on WordNet, it is necessary to properly cite the source. Citation figures are critical to WordNet funding.

About WordNet

Due to funding and staffing issues, we are no longer able to accept comment and suggestions.

We get numerous questions regarding topics that are addressed on our FAQ page. If you have a problem or question regarding

Fig 1. The snapshot of lexical database for english (WordNet)



WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: renal failure [Search WordNet]

Display Options: (Select option to change) [Change]
Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **renal failure**, kidney failure (inability of the kidneys to excrete wastes and to help maintain the electrolyte balance)
    - *direct hyponym* / *full hyponym*
    - *direct hypernym* / ***inherited hypernym*** / *sister term*
        - S: (n) kidney disease, renal disorder, nephropathy, nephrosis (a disease affecting the kidneys)
            - S: (n) uropathy (any pathology of the urinary tract)
                - S: (n) pathology (any deviation from a healthy or normal condition)
                    - S: (n) ill health, unhealthiness, health problem (a state in which you are unable to function normally and without pain)
                        - S: (n) pathological state (a physical condition that is caused by disease)
                            - S: (n) physical condition, physiological state, physiological condition (the condition or state of the body or bodily functions)
                                - S: (n) condition, status (a state at a particular time) *"a condition (or state) of disrepair"; "the current status of the*

Fig 2. The snapshot of WordNet content for renal failure

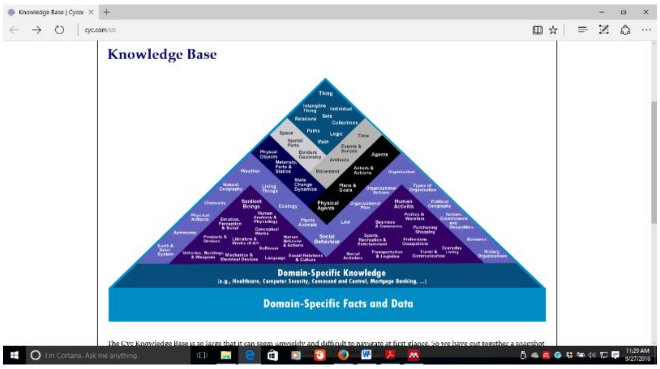
The hyponym/hypernym relationship (i.e., is-a relationship), and the meronym/holonym relationship (i.e., part-of relationship) are the most recognized relationships in WordNet. WordNet also introduces a larger amount of abstract concepts at the top of the taxonomic tree [13]. This is due to it being a general lexical database that does not merely focus on a singular domain. Figure 1 denotes the snapshot of WordNet web pages. The WordNet typically displays information such as synonym (S), direct hyponym (children of concept), direct hypernym (direct parents), full hy-ponym (all children), inherited hypernym (all parents), and sister term (shared direct parents). The WordNet contains a description of the concept in the form of tree structure as displayed in Figure 2.

SENSUS

SENSUS is an ontology that has 90,000 concepts of terminology taxonomy. Additional knowledge can also be placed in SENSUS [12]. SENSUS is an extension and reorganization of the WordNet. The added concept is realised at the top level of the Penman Upper Model, additionally to the rearrangement of the major branches of the Word-Net. Each concept in SENSUS is represented by one node, where each word has a unique specific sense, and the concepts are linked to is-a hierarchy. SENSUS can be browsed using the viewer Ontosaurus (http://mozart.isi.edu:8003/sensus2).

Cyc knowledge base (Cyc KB)

Cyc KB is a knowledge base designed to serve as an encyclopaedic repository of all human knowledge, pri-marily knowledge on common sense. Cyc KB is composed of terms and assertions relating to those terms. Funda-mental human knowledge can be included in Cyc KB such as facts, rules of thumb, and heuristics for reasoning about the objects and events of everyday life. At the present time, the Cyc KB contains over five hundred thousand terms, including seventeen thousand types of relations, additionally to seven million assertions which relate these terms [12], [14]. New assertions are continually added to the knowledge base through a combination of automated and manual means. Many more concepts can be expressed functionally, thereby enabling the automatic creation of millions of non-atomic terms, such as LiquidFn Nitrogen being used to describe liquid nitrogen. Additionally, the Cyc KB adds a vast number of assertions to the knowledge base (KB) by itself as a result of the inferencing process. Cyc can be browsed using http://www.cyc.com/kb/. Figure 3 shows a snapshot of the Cyc KB web page.



Knowledge Base

The Cyc Knowledge Base is so large that it can seem unwieldy and difficult to navigate at first glance. So we have put together a snapshot

Fig 3. The snapshot of the Cyc web page.

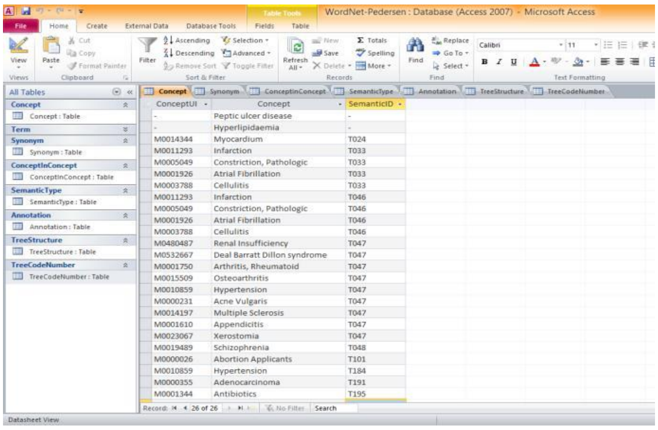
Datasets that used in this research is Wordnet. The reason evaluation between two terms in Wordnet because the WordNet has a larger amount of abstract concepts at the top of the taxonomic tree such as entity, abstract entity, abstraction, attribute, state, condition until to diabetis mellitus for concept Type 1 Diabetis. This is due to the intrin-sic nature of the WordNet as a general lexical database that not only focuses on one domain.

III. Methodology

A methodology is a set of ideas or guidelines about how to proceed in gathering and validating knowledge of a subject matter. To ensure the effectiveness of the system in the future, all aspects should be em-phasized. This research is focusing on X-Similarity feature-based method. This research was created by using WordNet dataset. This research has four phases. The first phase is focusing on data preparation. Second phase is similarity measure. Third phase is feature based method approach and the last phase is analysis result.

A. Data Preparation

There are a lot of data sources in semantic similarity such as Chemical Entities Biological Interest (CheBi), Gene Ontology (GO), National Centre Biological Ontology (NCBO) and others. The dataset that have been chosen in this research are chemical dataset which are taken from WordNet. Dataset of WordNet consisted of term, synonym, concepts and others. This research had used benchmark from WordNet and only use one dataset to make compari-son based on x-similarity featured based method. Figure 4 shows the WordNet datasets.



ConceptUI	Concept	SemanticID
-	Peptic ulcer disease	-
-	Hyperlipidaemia	-
M0014344	Myocardium	T024
M0011293	Infarction	T033
M0005049	Constriction, Pathologic	T033
M0001926	Atrial Fibrillation	T033
M0003788	Cellulitis	T033
M0011293	Infarction	T046
M0005049	Constriction, Pathologic	T046
M0001926	Atrial Fibrillation	T046
M0003788	Cellulitis	T046
M0480487	Renal Insufficiency	T047
M0532667	Deal Barratt Dillon syndrome	T047
M0001750	Arthritis, Rheumatoid	T047
M0015509	Osteoarthritis	T047
M0010859	Hypertension	T047
M0000231	Acne Vulgaris	T047
M0014197	Multiple Sclerosis	T047
M0001610	Appendicitis	T047
M0023067	Xerostomia	T047
M0019489	Schizophrenia	T048
M0000026	Abortion Applicants	T101
M0010859	Hypertension	T184
M0000355	Adenocarcinoma	T191
M0001344	Antibiotics	T195

Fig. 4 Wordnet dataset

B. Similarity Measure

Since we have to measure the data in semantic similarity, we have to choose the technique that will be use. There is a lot of technique. For example, information content, feature based technique and others.

In this research, we will use feature based approach technique. We have to test two datasets in one method. Figure 5 shows the Phase 2.

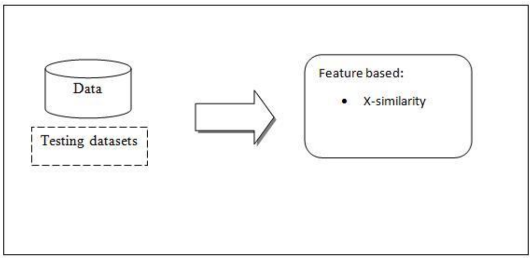


Fig 5. Process in second phase

C. Feature Based Process

This research had selected WordNet dataset, while in pre-processing; this research chooses thirty data from the dataset. Then we make comparison between the data in similarity process. If the data similar, then we will make calculation based on the formula each method. Based on the phase 1, the method is picking and constructed based on existed algorithm. The data test in X-Similarity method (2006).

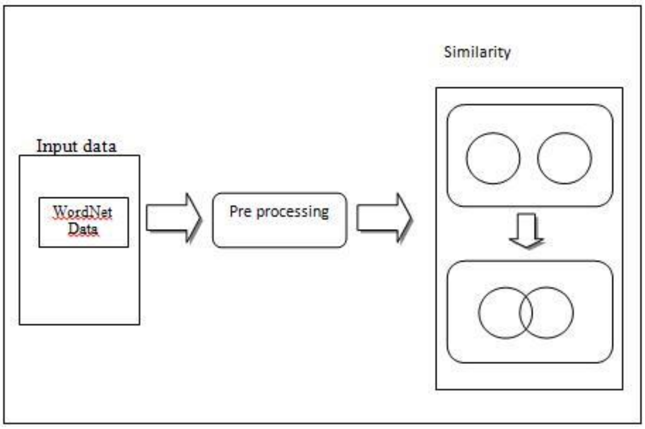


Fig. 6 Process in phase 3

D. Analysis the Result

Based on the Figure 6 phase 3, after testing the data, we have to compare the results. The value that will test is between 0 and 1 which means 0 is not similarity while 1 is exactly similar. If the value is more than 0, then the value will approximate to 1. If the value is not equal to 0, calculation with neighbourhood and description similarity will take place. Then the value with the highest correlation will be taken.

IV. Simulation Results and Analysis Abstract Requirements

This section describes the results and analysis for this research. The interface is created using Adobe Dreamweaver CS3 while programming language is PHP.

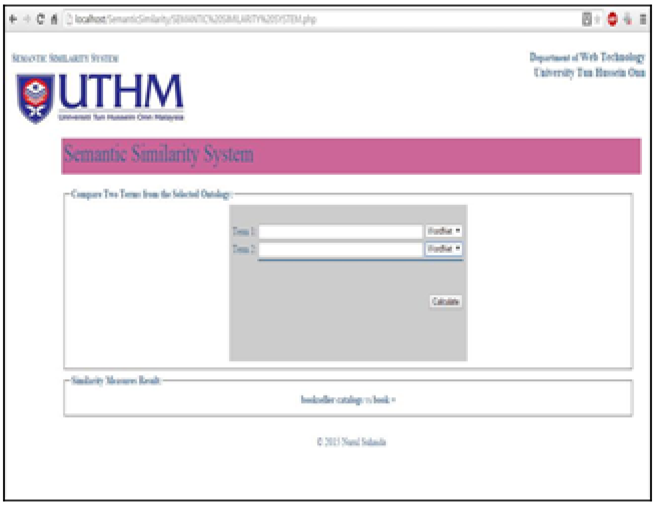

localhost/SemanticSimilarity/SEMANTIC%20SIMILARITY%20SYSTEM.php
Semantic Similarity System
Department of Web Technology
University Tun Hussein Onn
UTHM
Universiti Tun Hussein Onn Malaysia
Semantic Similarity System
Compare Two Terms from the Selected Ontology
Term 1:
Term 2:
WordNet
WordNet
Calculate
Similarity Measures Result
bookseller catalog vs book =
© 2015 Nurul Suhaida

Fig. 7. Interface

Based on semantic relations, two terms are similar if the concepts of the words and the concepts in their neighbourhood are lexically similar. Let A and B be two synsets or term description sets. It is because not all the term presents a connection with the same relationship (SR), type, example, Is-A and Part-Of [3]. The proposed similarity measure is expressed as follows:

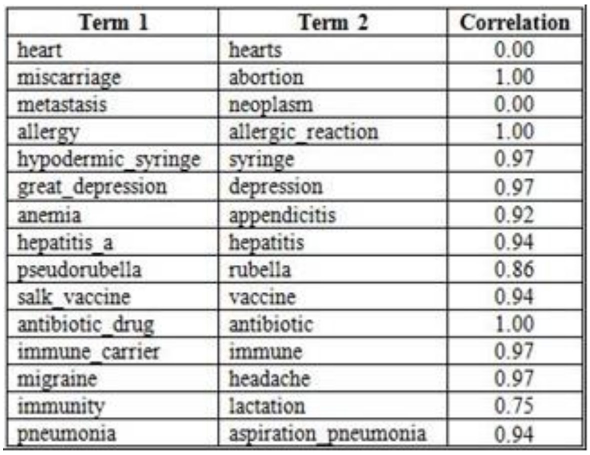
$$S_{synset}(a,b)=\frac{|A\cap B|}{|A\cup B|}\ if\ S_{synsets}(a,b)>0$$

$$S_{neighborhood}(a,b)=\frac{|A_i\cap B_i|}{|A_i\cup B_i|}\ S_{description}(a,b)=\frac{|A_i\cap B_i|}{|A_i\cup B_i|}\ if\ S_{synsets}(a,b)<0 \quad (1.1)$$

$$Sim(a,b)=\left\{\frac{1,}{\max\{S_{neighborhood},S_{description}\}}\right\} \quad (1.2)$$

Equation (1.2), is taken after taken the result from equation (1.1). From the equation (1.2), when the result is more than 0, then assume the result as a 1. If the result is equal to 0 then calculate the data by using $S_{neighbourhood}$ and S_{descr}. After that, compare the value from $S_{neighbourhood}$ and S_{decr}. Take the highest value which means if Sneighbourhood value is highest than S_{descr}, then take the value from $S_{neighbourhood}$.

Table 1. Result Corelation



Term 1	Term 2	Correlation
heart	hearts	0.00
miscarriage	abortion	1.00
metastasis	neoplasm	0.00
allergy	allergic_reaction	1.00
hypodermic_syringe	syringe	0.97
great_depression	depression	0.97
anemia	appendicitis	0.92
hepatitis_a	hepatitis	0.94
pseudorubella	rubella	0.86
salk_vaccine	vaccine	0.94
antibiotic_drug	antibiotic	1.00
immune_carrier	immune	0.97
migraine	headache	0.97
immunity	lactation	0.75
pneumonia	aspiration_pneumonia	0.94

CONCLUSION

This chapter concludes the use of information in this research. This research requires the information and knowledge in PHP language to be used. This research helps the user or programmers to gain something new and also can share their knowledge in term of semantic web and semantic similarity. This research can be updated by adding new dataset that different in this research or by using the same datasets with the benchmark on it.

This research had discussed about how the semantic similarity works. In addition, it also discuss from the beginning of this research until the end process of this research. Besides that, this research also explains about the method and purpose of making this research. By the end of chapter, user can implement feature based method and meas-ure the datasets in X-Similarity (2006) feature method

REFERENCES

[1] SEMILAR: A Semantic Similarity Toolkit (http://deeptutor2.memphis.edu/Semilar-Web.

[2] A. Goni, E. Mena and A Illarramendi, “Querying Heterogenous and Distributed Data Repositories Using Ontologies,” Information Modelling and Knowledge Base

[3] Petrakis, E. G. M., Varelas, G., Hliaoutakis, A., & Raftopoulou, P. (2006) “X-Similarity:Computing Similarty between Concepts from Different Ontologies”.

[4] Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering: Principles and methods. Data & Knowledge Engineering, 25(1–2), 161–197. https://doi.org/10.1016/S0169-023X(97)00056-6.

[5] Spasic, I., Ananiadou, S., McNaught, J., & Kumar, A. (2005). Text mining and ontologies in biomedicine: Making sense of raw text. Briefings in Bioinformatics, 6, 239–251. https://doi.org/10.1093/bib/6.3.239.

[6] Al-Mubaid, H., & Nguyen, H. A (2009). Measuring semantic similarity between biomedical concepts within multiple ontologies. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, 39(4), 389–398. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5061528

[7] Hliaoutakis, A. (2005). Semantic Similarity Measures in MeSH Ontology and their application to Information Retrieval on Medline. Technical University of Crete, Greek, Thesis PhD.

[8] Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38, 39–41. https://doi.org/10.1145/219717.219748

[9] Noy, N., & McGuinness, D. (2001). Ontology development 101: A guide to creating your first ontology. Development, 32, 1–25. https://doi.org/10.1016/j.artmed.2004.01.014.

[10] Choi, N., Song, I.-Y., & Han, H. (2006). A Survey on Ontology Mapping, College of Information Science and Technology Drexel University, Philadelphia, PA 19014 35(3): 34–41.

[11] Saruladha, K., Aghila, G., & Bhuvaneswary, A. (2011). Information content based semantic similarity approaches for multiple biomedical ontologies. Communications in Computer and Information Science (Vol. 191 CCIS, pp. 327–336). https://doi.org/10.1007/978-3-642-22714-1_34.

[12] Slimani, T. (2013). Description and Evaluation of Semantic similarity Measures Approaches. Journal of Computer Applicationsof Computer Applications, 80–no 10, 1–10.

[13] Solé-Ribalta, A., Sánchez, D., Batet, M., & Serratosa, F. (2014). Towards the estimation of feature-based semantic similarity using multiple ontologies. Knowledge-Based Systems, 55, 101–113. https://doi.org/10.1016/j.knosys.2013.10.015

[14] Singh, P. (2004). Web Ontology to Facilitate Semantic Web. 2nd International CALIBER-2004, New Delhi, 11-13 February, 2004 (pp. 11–13).