Interlingua: Linking Textbooks Across Different Languages

Isaac Alpizar-Chacon[0000−0002−6931−9787] and
Sergey Sosnovsky[0000−0001−8023−1770]

Utrecht University, Utrecht, The Netherlands
{i.alpizarchacon, s.a.sosnovsky}@uu.nl

Abstract. Increasing numbers of students enrol in formal and informal courses taught in a foreign language. Studying a course from an unfamiliar university/program is difficult enough, but the difficulties multiply when the transition to new course requirements is exacerbated by the necessity to learn course material in a foreign language. This paper describes Interlingua a platform where students can study textbooks in a foreign language supported by on-demand access to relevant reading material in their mother tongue. Interlingua automatically recognises important terminology within textbooks content, extracts structural models of textbooks and links sections and subsections across textbooks in different languages covering the same academic subject. The interface and architecture of Interlingua as well as the technologies underlying the platform are described.

Keywords: Linking textbooks · Modelling textbooks · Terminology extraction.

1 Introduction

1.1 ”Inter-lingual” students

Two parallel trends exist in the current EU education, independent one from another, originating from different conditions, yet leading to a shared outcome. From the socio-economic perspective, EU promotes ever-increasing mobility, especially when it comes to younger population. The Bologna process [6], the Youth on Move initiative [10], and students exchange programs like Erasmus contribute to the vision of a joint European education ecosystem, where students from all EU countries freely and actively engage in educational programs and individual courses across borders and cultures.

From the pedagogical (and technological) perspective, new forms of learning have emerged supported by information and communication technologies. They facilitate free and easy access to learning materials and promote more central and active role of a learner. Initiatives like Open Educational Resources (OER) and phenomena like Massive Open Online Course (MOOC) shape the new educational reality where students have more choice and flexibility in terms of which textbook to read, which course to take, and which skill to acquire. As a result they become less dependent on the actual institution issuing a degree.

These two trends reinforce each other and jointly contribute to a much-desired outcome of more scalable, sustainable and affordable education. However, they also lead to a potential problematic situation that is occurring more often as more international students enrol in formal university courses and/or free MOOCs taught in a foreign language. Studying a course from an unfamiliar university/program is challenging enough. It might be taught on a new (more abstract or intensive) level. It might require a student to have prerequisite knowledge and skills that s/he has not acquired yet. Yet for foreign students, this transition to new course requirements is aggravated by the necessity to learn material in a foreign language that they did not use when taking the prerequisite courses. A foreign student inevitably faces a certain language barrier amplified by the mismatch in the background knowledge and terminology. Unfortunately, the current tradition of resolving these difficulties is hardly efficient - teachers report that non-native students are allowed to bring dictionaries to regular classes and exams. As a result, we have an educational system that promotes student mobility on the level of policies, but does not sufficiently support it in on the individual level.

An effective remedy to this problem is the provision of international students with multilingual access to instructional material, where educational resources in a language of a course are accompanied by resources in their native language. There are two principle ways to achieve this: translation of the original resource and linking between two corresponding ”inter-lingual” resources. The translation-based approach will not help solve the problem: manual translation is not scalable, and machine-based translation is not yet capable of producing results of adequate quality in a narrow academic domain. In this paper, we present a solution based on automated semantic linking of related educational resources across languages with the main focus on textbooks[1]. Probability theory and statistics has been chosen as the target domain. One reason for it is the popularity of this subject in teaching programs of many technical and social science degrees. Another reason is that this subject uses a lot of specific terminology both new and borrowed from prerequisite parts of mathematics.

1.2 Textbooks modelling and linking

Textbooks are often considered non-structured information resources for the purpose of text analysis and information extraction. Yet, textbooks are created, structured and formatted by their authors (who are presumably domain experts) with a primary purpose to explain the knowledge in the domain to a novice. A textbook author uses his/her own understanding of the domain when structuring and formatting the content of the textbook to facilitate this explanation. As a result, the formatting and structural elements of a textbook (headers, table of

[1] This work was supported by the INTERREG-IVA-GR program (grant 138 GR DeLux 32274)

content (ToC), index) not only shape the organisation of the textbook, but also reflect the organisation of the domain as the author sees it. When extracted they can be formally represented as a semantic model of the textbook itself and the domain it teaches. We have experimented with automated extraction of such models from multiple textbooks in different domains (probability theory and statistics and information retrieval) and different representation formats (PDF and ePUB) [1].

A textbook model can support a variety of tasks including inference and reasoning about knowledge taught bu the textbook, tracking and modelling of reader's progress with the textbook, enhanced interaction through semantic search and adaptive navigation through the textbook. In the case of multiple textbooks in the same domain, such models can be mapped between each other providing a semantic bridge for linking related sections and fragments across textbooks. An external reference model can be used to facilitate the mapping and help resolving possible terminology conflicts. In [2], we report the results of several experiments on mapping automatically extracted textbook models between each other and to DBPedia [4]. An alternative to DBPedia can be another global knowledge graph available in the Open Linked Data Cloud, such as Freebase [5], Wikidata [21], YAGO [19] or any other encyclopedic data set. Another alternative can be a domain-oriented ontology or a thesaurus. In either case, to support linking across textbooks in different languages, a reference model has to specify terminology in these languages. In this project, we have used two such multilingual models: DBPedia (which extracts textual labels and descriptions of entities in multiple languages from corresponding resources of Wikipedia) and the multilingual glossary of statistical terms maintained by the International Statistical Institute (ISI) - we will refer to it as the ISI glossary from now on.

The rest of this paper is structured as follows. Section 2 gives an overview of related work in the field of textbook model extraction and linking. Section 3 presents the details of the proposed technology and the architecture of Interlingua - the Web-based platform that students can use to access textbooks in a foreign language while getting recommendations of related reading resources in their mother tongue. Section 4 presents the interface of Interlingua and the ways it can support students. Section 5 concludes the paper with a discussion and a description of future work.

2 Related work

Creation of knowledge models from textbooks has been explored in a limited way. Larrñaga et al. [14] used natural language processing techniques, heuristic reasoning, and ontologies to semi-automatically construct a representation of the knowledge to be learned from electronic textbooks. Their approach uses the document outline to create a tree-like internal representation, and extract the main domain topics and the pedagogical relationships among them. Sosnovsky et al. experimented with harvesting topic-based models from HTML textbooks based on the structure of their headings; the resulting models have been automatically mapped into a reference ontology to facilitate more fine-grained inference and adaptation [18]. Wang et al. [24] have extracted concept hierarchies from textbooks using Wikipedia. Textbooks sections were matched against Wikipedia articles and, as a results, annotated with corresponding Wikipedia entities. The hierarchy was reconstructed based on the hierarchy of chapters, sections and subsections. Olney et al. [17] used a combination of natural language processing techniques and a manually extracted index to generate concept maps from a biology textbook.

Several projects experimented with semantic linking of relevant textbooks written in the same language. Guerra et al. [12] used a probabilistic topic modelling approach to extract topic models from textbook and use them to link sections and subsections across multiple textbooks. Later, Meng et. al [16] explored different content modelling approaches for textbook linking: a term-based approach (each word is considered as a knowledge component), LDA (latent topics representing knowledge components), and a concept-based approach (author-assigned keywords in scientific publications representing knowledge components). Also, an ensemble of the three approaches was used. The semantic and the combined approaches achieved valuable linking performance.

3 Interlingua technology

Architecturally, Interlingua consists of two large components. The offline component performs the tasks of textbook modelling and linking, while the online component supports students' interaction with the content of linked textbooks. This section describes the details of the offline component; the student interface is presented in the next section.

The overall process of adding a new textbook into the Interlingua content repository is depicted by 1. The only action performed manually is the upload of textbook files. A teacher decides which textbooks and in which languages should be available for the students of the course. After a textbook file is submitted, the textbook model is generated: the textbook is divided into sections and sub-sections, its ToC and index are extracted, each page is identified, every differently formatted fragment is recognised and provided with a semantic label if a corresponding rule exists (we keep working on expanding the rule base). Index plays a special role in this process, as it provides a glossary of manually selected terms that the author of the textbook deemed meaningful. Interlingua extracts index terms and pages referenced by the terms and uses them as the semantic anchors to link pages and sections of the textbook to the concepts of the reference ontology and through them to other textbooks available in the content repository. Finally, the self-assessment component uses the information from DBPedia to generate multiple choice questions related to the index terms explained by the current section of the textbook. Once processed and stored this way, the textbook becomes accessible through the student interface. The rest of the section describes individual phases of this process; however, we refer readers to our two other papers for a detailed description of textbook modelling [1] and linking [2].

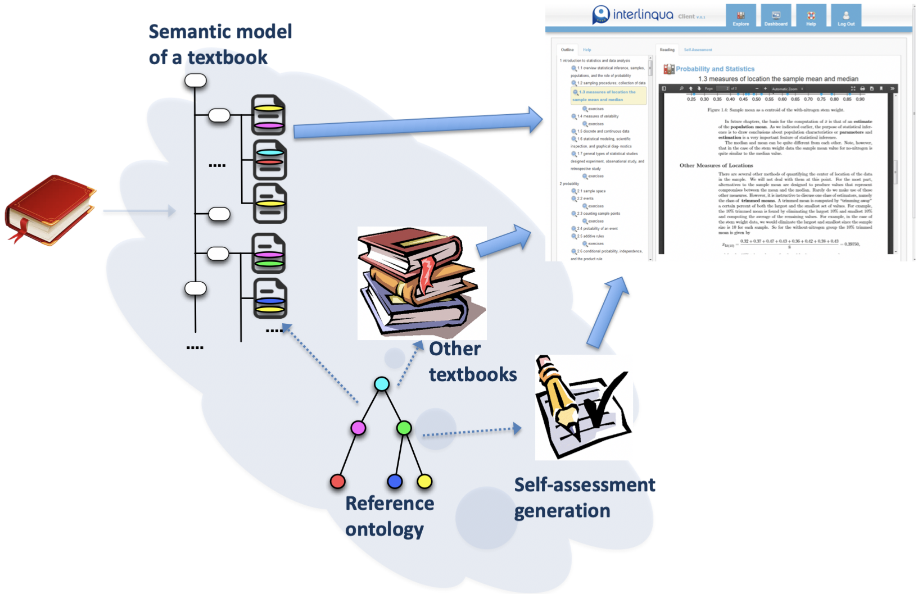

Semantic model
of a textbook
....
....
....
Other
textbooks
Reference
ontology
Self-assessment
generation
interlinqua Client v.0.1
Explore
Dashboard
Help
Log Out
Outline
Help
Reading
Self-Assessment
1 introduction to statistics and data analysis
1.1 overview statistical inference, samples, populations, and the role of probability
1.2 sampling procedures; collection of data
1.3 measures of location the sample mean and median
exercises
1.4 measures of variability
exercises
1.5 discrete and continuous data
1.6 statistical modeling, scientific inspection, and graphical diag- nostics
1.7 general types of statistical studies designed experiment, observational study, and retrospective study
exercises
2 probability
2.1 sample space
2.2 events
exercises
2.3 counting sample points
exercises
2.4 probability of an event
2.5 additive rules
exercises
2.6 conditional probability, independence, and the product rule
Probability and Statistics
1.3 measures of location the sample mean and median
Page 2 of 3
Automatic Zoom
0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90

Figure 1.4: Sample mean as a centroid of the with-nitrogen stem weight.

In future chapters, the basis for the computation of $\bar{x}$ is that of an **estimate** of the **population mean**. As we indicated earlier, the purpose of statistical inference is to draw conclusions about population characteristics or **parameters** and **estimation** is a very important feature of statistical inference.

The median and mean can be quite different from each other. Note, however, that in the case of the stem weight data the sample mean value for no-nitrogen is quite similar to the median value.

Other Measures of Locations

There are several other methods of quantifying the center of location of the data in the sample. We will not deal with them at this point. For the most part, alternatives to the sample mean are designed to produce values that represent compromises between the mean and the median. Rarely do we make use of these other measures. However, it is instructive to discuss one class of estimators, namely the class of **trimmed means**. A trimmed mean is computed by "trimming away" a certain percent of both the largest and the smallest set of values. For example, the 10% trimmed mean is found by eliminating the largest 10% and smallest 10% and computing the average of the remaining values. For example, in the case of the stem weight data, we would eliminate the largest and smallest since the sample size is 10 for each sample. So for the without-nitrogen group the 10% trimmed mean is given by

$$\bar{x}_{tr(10)} = \frac{0.32 + 0.37 + 0.47 + 0.43 + 0.36 + 0.42 + 0.38 + 0.43}{8} = 0.39750,$$

Fig. 1. Overall process of adding a textbook into Interlingua

3.1 Textbook modelling

The Textbook modelling component is in charge of processing new textbooks and creating their internal representations. The process of extraction of semantic models from PDF textbooks consists of four steps: extraction of raw content, construction of the style library, identification of logical elements, and construction of the model. Each step and its tasks are depicted in Fig. 2. Each of these four main steps starts after the previous and relies on its results. We focus mainly on PDF as the most common representation format; however, the overall approach is also applicable to other formats that are more explicit and coherent in their structural specifications than PDF.

Raw content extraction. In this step, the PDF file of the textbook is parsed to extract the text (the characters), geometrical (X- and Y- coordinates on the pages) information and formatting (type, size, and features of fonts) styles of the document. The parsing of the PDF is done with the Apache PDFBox library[2].

[2] https://pdfbox.apache.org/

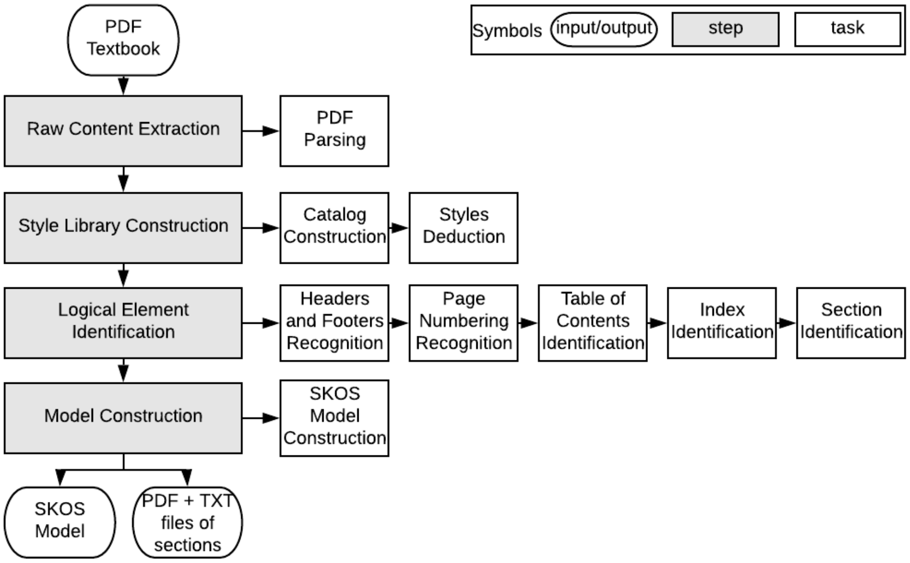


Fig. 2. Steps and tasks of the textbook modelling component

The parser reads the PDF content stream where the elements are sorted from left to right and top to bottom according to their position on a page. It merges smaller objects into bigger using a bottom-up approach. First, individual characters are processed. Then, characters are grouped into words, words into lines, and lines into pages. At each stage, the extractor compares the proximity of an object's coordinates with the coordinates' of its preceding neighbour to form a bigger object. Besides, the content itself, the parser also saves the styling information if every object, such as font size, family, properties (boldness, colour, etc.) as well as margins.

Style library construction. Once all text fragments are extracted and their styles are identified, the library of all the formatting styles used in the textbook is constructed. This library is used in the next step to recognise the different structural elements constituting the textbook. The main style of the textbook content is identified, as well as styles for important text fragments, headings of sections and subsections, etc. Geometric features of the text play a role in this process as well, as they allow to combine text into paragraphs, sections and chapters, allow to compute spacing and indentations, allow to recognise columns and captions, etc.

Logical element identification In this step, some auxiliary texts such as headers, footers and page numbers are recognised and separated from the main content. In addition, important structural elements of the textbook are identified. The ToC section is recognised and parsed to construct an outline of the textbook. Each content section and sub-section is identified, associated with its header, its ToC entry, and its page interval. Finally, the textbook index is processed. This section plays a special role, as it provides the main source of information for subsequent textbook linking. A good textbook index is not just a collection of words, but, essentially, a reference model produced by a domain expert according to a predefined set of rules. Every publisher guides the process of index creation stipulating index length and style, suggesting what can be good and bad candidates for index terms, advising on how to maintain consistency when creating hierarchical indices, etc. [3]. Each index entry is provided with one or more links to the pages within a textbook. And these pages do not simply mention the entries, but provide meaningful references by either introducing corresponding terms or elaborating them.

Model construction Finally, with all the extracted information a semantic knowledge model of the textbook is constructed. The model allows to quickly obtained different knowledge views of the textbooks to provide rich information to the users, for example: hierarchical structure of sections, all index terms related to one section, all pages where an index term is explained, etc. The model is represented using SKOS[3], which is a W3C standard based on RDF and OWL. It provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies and other similar types of controlled vocabulary.

3.2 Textbook enrichment

This component takes the model of a textbook and links it to two external reference models: the ISI Glossary and DBpedia. The ISI Glossary[4] was created by The International Statistical Institute. It contains more than 3500 statistical terms combined into synsets and translated into 31 different languages. The glossary has been parsed and represented as a reference ontology. As a result, Interlingua can unambiguously recognise that, for example, in English term ”z-score” is the same as ”standard score”. What is even more importantly, we can also identify all the ways this term is called in all 31 languages of the ISI glossary. DBpedia is a machine-understandable source of knowledge created based on automatically-extracted structured information from Wikipedia, which contains 4.58 million resources and is available in 125 languages. It plays two important roles in Interlingua. First, it provides and alternative domain-independent multilingual reference model. Although, ISI glossary has a much better granularity and coverage statistical terms, DBPedia ensures that the overall approach of Interlingua is applicable to other domains (and other languages) where a dedicated high-quality multilingual thesauri cannot be obtained. The second role of DBPedia is to provide additional structural and annotation information, such as

[3] https://www.w3.org/TR/skos-reference/

[4] http://isi.cbs.nl/glossary/

dbpedia-owl:abstract for term explanations, dcterms:subject for fields of study where the term is coming from and dbpedia-owl:wikiPageWikiLink for different terms, which are mentioned in the same context, etc. The enrichment process is shown in Fig.3. More information about the enrichment algorithm is available in [2]

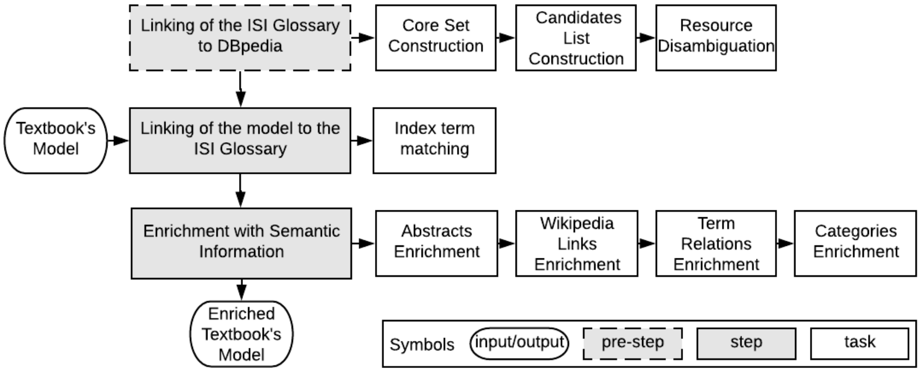

Linking of the ISI Glossary to DBpedia
Core Set Construction
Candidates List Construction
Resource Disambiguation
Textbook's Model
Linking of the model to the ISI Glossary
Index term matching
Enrichment with Semantic Information
Abstracts Enrichment
Wikipedia Links Enrichment
Term Relations Enrichment
Categories Enrichment
Enriched Textbook's Model
Symbols
input/output
pre-step
step
task

Fig. 3. Enrichment approach

Linking of the ISI glossary to DBpedia. In the beginning terms of the ISI Glossary have been (automatically) linked to its correspondent entities in DBpedia. The linking algorithm first constructs a core set of terms from the glossary for which a matching DBpedia resource was unambiguously discovered. Then, it recursively uses textual content associated with recognised DBPedia entities as the context information to facilitate disambiguation of newly discovered candidate matches. The process repeats itself until no new terms can be match with sufficient certainty. This only needs to be done once, before the enrichment of individual textbook models.

Linking of the textbook model to the ISI Glossary. At this step, each index term extracted from the textbook is compared against each term of the ISI Glossary, and when the similarity is over 90 %, the index term is linked to the ISI Glossary term.

Enrichment with semantic information. Finally, for each index term with a linked ISI Glossary term that has a DBpedia resource, semantic information is extracted to enrich the model of the textbook. The following information is obtained from the linked DBpedia resources: abstracts, links to the corresponding Wikipedia pages of the resources, DBpedia categories of the resources, and direct relations among the linked index terms using the links that exist in the

Wikipedia pages of the resources. At the end of this approach, the textbook's model is updated with the gathered information.

3.3 Textbook linking

The linking among the different parts of textbooks is done based on the ISI glossary as a reference model to build a Vector Space Model (VSM). The VSM consists of several documents (each section) per textbook in the rows, and each entry of the ISI glossary as the columns. The weight for each term of each document is calculated from the strength of the association between the term and the document. The VSM is improved pruning the terms in the glossary that are not found in the corpora. After the VSM has been created, it is split into matrices, one for each language, and a similarity matrix is created for each language pair by first multiplying the two appropriate parts of the VSM together and then normalising the resulting matrix. These matrices contain the similarity value for any pair of sections of textbooks from different languages. Finally, each section of each textbook is linked to the five most similar documents in each target textbook to create the links among the textbooks in different languages.

3.4 Assessment generation

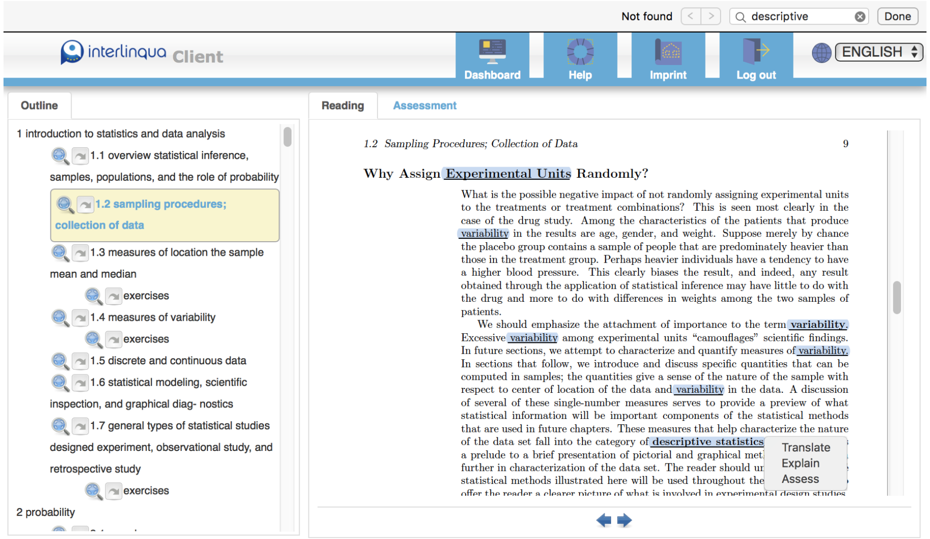
The assessment engine generates Multiple Choice Questions (MCQ) for the learners to examine their understanding of the terminologies related to the section that they are currently reading. The basic MCQ asks the learners to select the correct translation of one of the introduced concepts in the current section into their native language. Each question has only one right answer and several distractors. To generate the questions the assessment engine uses the ISI glossary to get the labels in different languages for the same concept. Distractors are chosen using the relations among concepts obtained from the enrichment of the model to select reasonably difficult distractors. In other words, the semantic distance between the correct answer and a distarctor in the model should not be very long and the lengths of both textual labels should be comparable.

4 Interlingua interface

When a student accesses the Interlingua client, the entry page shows a list of available textbooks in the target language of study. A student can indicate her language of study and the mother tongue. Currently, Intelingua uses the following textbooks: [8, 23] in English, [13, 20] in French, [9, 11] in German, [22] in Spanish, and [7, 15] in Dutch.

A student selects one of the available textbooks to open the textbook navigation page (see Fig. 4). It consists of the outline panel on the left side and the content panel with multiple tabs on the right side of the window.

In the outline panel, a student can brows through sections and sub-sections of the textbook. She can click on a magnifying glass icon annotating each section


Not found
descriptive
Done
interlinqua Client
Dashboard
Help
Imprint
Log out
ENGLISH
Outline
1 introduction to statistics and data analysis
1.1 overview statistical inference, samples, populations, and the role of probability
1.2 sampling procedures; collection of data
1.3 measures of location the sample mean and median
exercises
1.4 measures of variability
exercises
1.5 discrete and continuous data
1.6 statistical modeling, scientific inspection, and graphical diag- nostics
1.7 general types of statistical studies designed experiment, observational study, and retrospective study
exercises
2 probability
Reading
Assessment
Translate
Explain
Assess

1.2 Sampling Procedures; Collection of Data 9

Why Assign Experimental Units Randomly?

What is the possible negative impact of not randomly assigning experimental units to the treatments or treatment combinations? This is seen most clearly in the case of the drug study. Among the characteristics of the patients that produce variability in the results are age, gender, and weight. Suppose merely by chance the placebo group contains a sample of people that are predominately heavier than those in the treatment group. Perhaps heavier individuals have a tendency to have a higher blood pressure. This clearly biases the result, and indeed, any result obtained through the application of statistical inference may have little to do with the drug and more to do with differences in weights among the two samples of patients.

We should emphasize the attachment of importance to the term variability. Excessive variability among experimental units "camouflages" scientific findings. In future sections, we attempt to characterize and quantify measures of variability. In sections that follow, we introduce and discuss specific quantities that can be computed in samples; the quantities give a sense of the nature of the sample with respect to center of location of the data and variability in the data. A discussion of several of these single-number measures serves to provide a preview of what statistical information will be important components of the statistical methods that are used in future chapters. These measures that help characterize the nature of the data set fall into the category of descriptive statistics

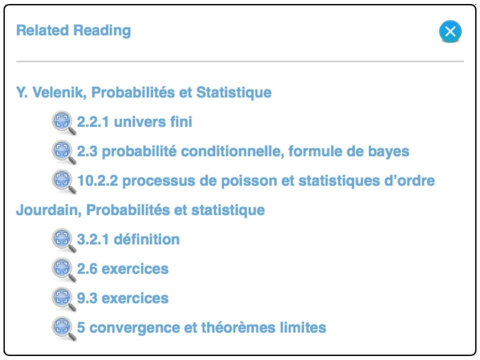
a prelude to a brief presentation of pictorial and graphical met

further in characterization of the data set. The reader should ur

statistical methods illustrated here will be used throughout the

offer the reader a clearer picture of what is involved in experimental design studies

Fig. 4. Textbook navigation page



Y. Velenik, Probabilités et Statistique

- 2.2.1 univers fini
- 2.3 probabilité conditionnelle, formule de bayes
- 10.2.2 processus de poisson et statistiques d'ordre

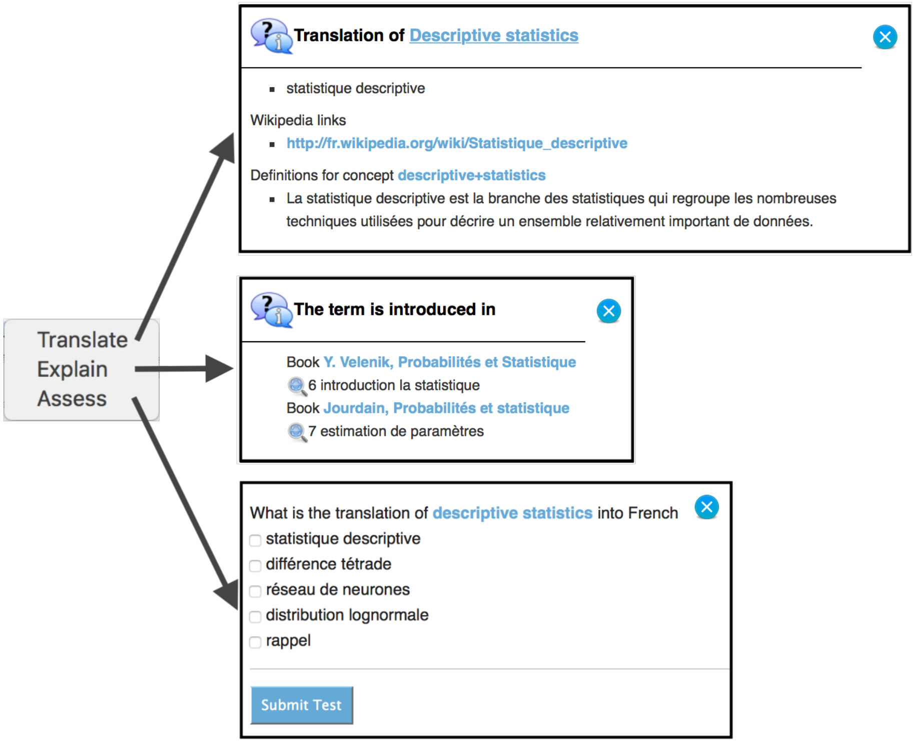
Jourdain, Probabilités et statistique

- 3.2.1 définition
- 2.6 exercices
- 9.3 exercices
- 5 convergence et théorèmes limites

Fig. 5. Related readings displayed in a pop-up

title to load the respective section into the content panel. The arrow icon is used to display a list of related readings of the selected sub-section in the mother tongue of the student (see Fig. 5).

The student can browse through a subsection content in a similar way to any standard PDF viewer application. Index terms are recognised and highlighted in the content to indicate that additional interaction with them is possible. When a student clicks in a highlighted word, an action menu appears with three available options: translate, explain, and assess (see Fig. 6). The *translate* action opens a pop-up window that presents the translation of the term into the mother tongue of the student, and the definition of the term extracted from DBpedia (if this index term has been found in DBpEdia). The *explain* action opens a pop-up window that gives access to the sections of the textbooks in the mother tongue of the user in which the term is explained. The *assess* action generates and displays in a pop-up window an question about the translation of the term into the mother tongue of the user.



Translation of Descriptive statistics

- statistique descriptive

Wikipedia links

- http://fr.wikipedia.org/wiki/Statistique_descriptive

Definitions for concept descriptive+statistics

- La statistique descriptive est la branche des statistiques qui regroupe les nombreuses techniques utilisées pour décrire un ensemble relativement important de données.

The term is introduced in

Book Y. Velenik, Probabilités et Statistique
6 introduction la statistique
Book Jourdain, Probabilités et statistique
7 estimation de paramètres

What is the translation of descriptive statistics into French

□ statistique descriptive
□ différence tétrade
□ réseau de neurones
□ distribution lognormale
□ rappel

Submit Test

Fig. 6. Action menu and pop-ups available for highlighted index terms

As mention before, in the navigation page a student also can open related readings for the current section in her mother tongue (see Fig. 4). A pop-up window suggests a list of related readings (see Fig. 5). if a student selects any of the suggested links, a corresponding sub-section will be loaded in a new tab. The student can switch between the reading tabs that now contain the related content in two different languages.

Finally, the user can click on the 'Assessment' tab in the content panel to generate an assessment composed of several MCQ related to the content of the currently browsed section (the questions are similar to the one at the bottom pop-up in Fig 6).

5 Discussion and future work

The paper has presented the motivation, the approach and the implementation of the Interlingua system that provides foreign students with on-demand access to related textbook material in different languages.

There might be several possible concerns both practical and pragmatic about the scale and applicability of this approach. One question that needs to be addressed is the availability of textbooks and the copyrights issues. In our experience, university libraries can supply enough PDF-based textbooks on a variety of subjects. From the point of copyright protection, if a system provides enhanced access to these books but only to the students of the university holding necessary subscriptions, then publishers do not have a reason to object. In the worst case scenario, many good-quality textbooks are freely available online nowadays in open repositories such as Openstax[5]/Connections[6], Open Textbook Library[7], OER-Commons[8], etc.

Another concern is how well the automated textbook modelling technology will cope with different textbook formatting. Naturally, formats can be quite different between different textbooks, yet the authors and the publishers make every effort to ensure they are consistent within textbooks (otherwise they would cause unnecessary confusion) and we believe, that the formatting patterns of textbooks do follow a set of logical rules. When implementing the rule-based component of the model extraction algorithm we try to capture these patterns. We have observed that the variability is quite manageable and as we have been processing more and more textbooks, we update our set of rules progressively less.

Finally, we realise that none of the extracted models is guaranteed to provide high-quality representation of a domain. These models can potentially suffer from several drawbacks: (1) Subjectivity: they can contain terms are only marginally related to an objective picture of the domain semantics; (2) Coverage: they can miss important terms if a textbook does not cover (enough) a particular part of a domain; (3) Granularity: the terms in an index can be too detailed, too broad, or inconsistently alternate in their granularity; (4) Lack of semantics: most indices are flat lists of terms with no relations between them. To combat these potential problems, glossaries extracted from individual textbooks can be integrated between each other and to external models: global as DBPedia, or domain-focused as ISI glossary. We plan to further investigate these and other issues. For example, an important direction for future work is to evaluate the effectiveness of the system in a user study with real students from a target group.

From a broader perspective, Interlingua is an interesting example of a service that can be built on top of linked semantic models extracted from related textbooks. However, a corpus of semantically linked high-quality educational

[5] https://openstax.org
[6] https://cnx.org/
[7] https://open.umn.edu/opentextbooks
[8] https://www.oercommons.org/

content can be used to implement a range of different services including adaptive navigation through or recommendation of textbook content, enrichment of textbook content with external (interactive) educational resources, or extraction of different types of learning objects from the textbook themselves.

Finally, it is interesting to explore thew applicability of the the approach towards automated modelling and linking of textbook that underlies Interlingua in other, less formal domains (e.g., medicine) or domains with conflicting viewpoints (e.g. history).

References

1. Alpizar-Chacon, I., Erensoy, O., Sosnovsky, S.: Order out of chaos: Construction of knowledge models from pdf textbooks. In: Proceedings of the 10th International Conference on Knowledge Capture (Submitted). K-CAP '19, ACM, New York, NY, USA (2019)
2. Alpizar-Chacon, I., Sosnovsky, S.: Expanding the web of knowledge: one textbook at a time. In: Proceedings of the 30th ACM Hypertext conference (Submitted). HT '19, ACM, New York, NY, USA (2019)
3. Ament, K.: Indexing: a nuts-and-bolts guide for technical writers. William Andrew (2001)
4. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The semantic web, pp. 722–735. Springer (2007)
5. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. pp. 1247–1250. AcM (2008)
6. Bologna Declaration: Towards the european higher european area (June 1999), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM:c11088
7. Caenepeel, S., de Groen, P.: Waarschijnlijkheidsrekening en statistiek. Vrije Universiteit Brussel (2002)
8. Dekking, F.M., Kraaikamp, C., P., L.H., Meester, L.E.: A modern introduction to probability and statistics: understanding why and how. Springer (2005)
9. Eckstein, P.P.: Repetitorium Statistik. Springer Gabler (2013)
10. European Commission: Youth on the move - an initiative to unleash the potential of young people to achieve smart, sustainable and inclusive growth in the european union (2010), http://europa.eu/youthonthemove/docs/communication/youth-on-the-move_EN.pdf
11. Fahrmeir, L.: Statistik: der Weg zur Datenanalyse. Springer (2007)
12. Guerra, J., Sosnovsky, S., Brusilovsky, P.: When one textbook is not enough: Linking multiple textbooks using probabilistic topic models. In: Hernández-Leo, D., Ley, T., Klamma, R., Harrer, A. (eds.) Scaling up Learning for Sustained Impact. pp. 125–138. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
13. Jourdain, B.: Probabilités et statistique (2013)
14. Larrañaga, M., Conde, A., Calvo, I., Elorriaga, J.A., Arruarte, A.: Automatic generation of the domain module from electronic textbooks: Method and validation. IEEE Trans. on Knowl. and Data Eng. **26**(1), 69–82 (Jan 2014)
15. Marchant, T.: Statistiek I. Universiteit Gent (2012)

16. Meng, R., Han, S., Huang, Y., He, D., Brusilovsky, P.: Knowledge-based content linking for online textbooks. In: 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI). pp. 18–25 (Oct 2016). https://doi.org/10.1109/WI.2016.0014
17. Olney, A., Cade, W., Williams, C.: Generating concept map exercises from textbooks. In: Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 111–119 (2011)
18. Sosnovsky, S., Hsiao, I.H., Brusilovsky, P.: Adaptation in the wild: ontology-based personalization of open-corpus learning material. In: European Conference on Technology Enhanced Learning. pp. 425–431. Springer (2012)
19. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th international conference on World Wide Web. pp. 697–706. ACM (2007)
20. Velenik, Y.: Probabilités et statistique. Universit de Genve (2012)
21. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledge base (2014)
22. Walpole, R.E.: Probabilidad y estadistica para ingenieros. Pearson (2012)
23. Walpole, R.E., Myers, R.H., Myers, S.L., Ye, K.: Probability statistics for engineers scientists. Prentice Hall (2012)
24. Wang, S., Liang, C., Wu, Z., Williams, K., Pursel, B., Brautigam, B., Saul, S., Williams, H., Bowen, K., Giles, C.L.: Concept hierarchy extraction from textbooks. In: Proceedings of the 2015 ACM Symposium on Document Engineering. pp. 147–156. ACM (2015)