Spoken Language Understanding (SLU) Module for Movie Domain using NL-SPARQL Data Set

|  |  |
| --- | --- |
| 000  00  00  00  00  00  00  00  00  00  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 | 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |

|  |
| --- |
|  |
| Rodrigo Joni Sestari  number: 179020  rodrigo.sestari@studenti.unitn.it |
|  |
|  |
|  |
|  |
|  |

Abstract

This document has the objective of illustrate the implementation of Develop Spoken Language Understanding (SLU) Module for Movie Domain given a NL-SPARQL Data Set. This document consists in the first project for the course of Language Understanding System of the University of Trento

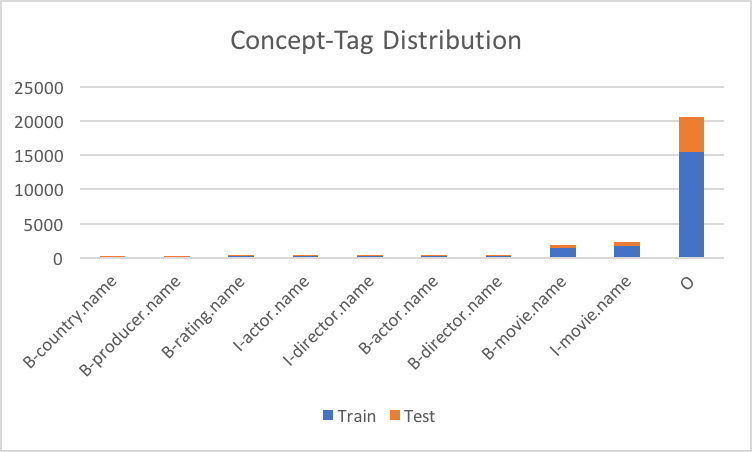
Introduction

There are some steps necessary to implement a SLU module for the Movie Domain. The first part consists in use the Train dataset in order to create a language model, de second one consist to apply this language model into a test dataset using different smoothing methods, the last part consist to evaluate these results.

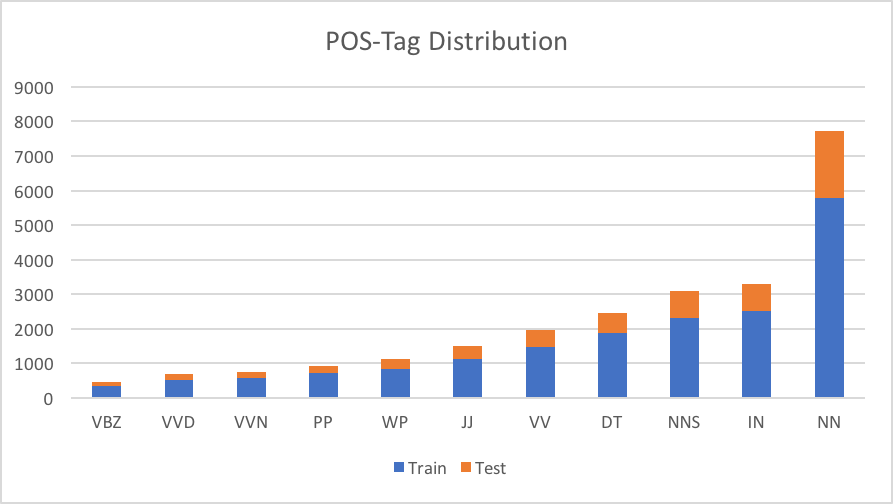
Data analisys

Manuscripts

Was given to datasets split by Train and test, the first one contains the Concept Tagging format by by the tuple <word, concept> the concept use IOB Notation. The second one contains additional features like Lemma and Pos-tags having the strcutre <word,pos-tag,lemma> the train dataset contains XXX words and XXX sentences , instead the test dataset contains xxx words an xxx sentendes



the additional dataset allows use lemmas, the use Lemmas is used to simplify the grammar making it less complex. in this paper the lemmas is not used.



Is possible see in the above graph the Pos-tag distribution

must be in two-column format. Exceptions to the two-column format include the title, authors’

Theory about natural language process

|  |  |
| --- | --- |
|  |  |

I this chapter introduces some principles about the natural language processing that come used in this paper:

* **N-gram**: n-gram is a contiguous sequence of n items from a given sequence of text or speech. In this paper the sequence come from the DataSet.
* **FSA**: is a mathematical model of computation. It is an abstract machine that can be in exactly one of a finite number of states at any given time. The FSA can change from one state to another in response to some external inputs; the change from one state to another is called a transition. A FSA is defined by a list of its states, its initial state, and the conditions for each transition. A Finite State Automaton (FSA) A consists of 4 objects - A set I called the input alphabet, of input symbols - A set S of states the automaton can be in; - A designated state s0 called the initial state; - A designated set of states called the set of accepting states,
* **FST**: A FST is a finite automaton whose state transitions are labeled with both input and output symbols. Therefore, a path through the transducer encodes a mapping from an input symbol sequence to an output symbol sequence
* **WFST**: A weighted transducer puts weights on transitions in addition to the input and output symbols. Weights may encode probabilities, durations, penalties, or any other quantity that accumulates along paths to compute the overall weight of mapping an input sequence to an output sequence. Weighted transducers are thus a natural choice to represent the probabilistic finite-state models prevalent in speech processing.
* **LM**: Language Model or ML compute the probability of a sentence or sequence of words: P(W) = P(w1,w2,w3,w4,w5...wn)
* **Cut-off**
* **Smoothing**: Assume that no n-gram of known words has 0 probability  and Redistribute probability mass from seen to unseen events, this is known as smoothing. In this work the methods which are used are Witten-Bell, Absolute, Katz, Kneser-Ney, Presmoothed and Unsmoothed

Tools

To execute the tasks about this paper was necessary the following tools:

* **OpenFST**: an open-source library for weighted finite-state transducers (WFSTs). OpenFst consists of a C++ template library with efficient WFST representations and over twenty-five oper- ations for constructing, combining, optimizing, and searching them.
* **OpenGRM**: is used for making and modifying n-gram language models encoded as weighted finite-state transducers (FSTs). It makes use of functionality in the OpenFst library to create, access and manipulate n-gram models. Operations for counting, smoothing, pruning, applying, and evaluating models as well as support for distributed computation are among those provided.
* **Conlleval**: A perl script that allow evaluate the SLU result.

Implementation

The implementation consist to execute a some script in order to train and test the LM, thesis script uses external tools like OpenFST and OpenGRM:

Scripts

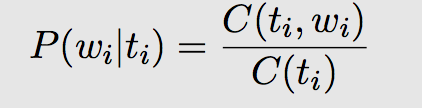
Some scripts are implemented in order to elaborate some data in input and produce formatted output files useful for the entire procedure. The scripts are:

* **Start.sh**: Used to start the Train/Test process, contains 3 parameters <smoothing=1..6, grammar=1..5, threshold=-1..2>. The parameter smoothing represent the smoothing method that can be (1=Absolute, 2=Katz,3=Kneser\_ney,4=Pre smoothed,5=unsmoothed,6=Witten\_bell The parameter grammar represent the Ngram order, the last parameter represents the Cut-off threshold, with 0 without cut-off. if threshold is negative, that script run all possibles combination of method N-gramar and threshold.
* **TrainTest.sh**: Script reposable to train and test the ML
* **transducer\_UNK.sh** : Create the Trasducer filer for Unknow tokens
* **transducer.sh**: Create the Trasducer filer for Unknow tokens

Train

Initially was created two files, POS.counts that represents C(ti) that contains the tuple <tag,count>, the second one called TOK\_POS.counts represent C(ti,wi) is represented by the triple <word,tag,count>

having these two files, the next step was create the file TOK\_POS.probs this file represents the probability of see the current word given the current tag P(wi|ti),



if C(ti,wi) <= threshold Cut-off then the probability is not inserted into TOK\_POS.probs

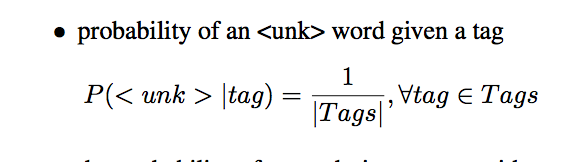
The last file to be create is the Lexico file, its contains the vocabulary used in the Train, The Lexico file was create using the file TOK\_POS.counts, in the head file was put <epsilon> 0 and at the end of file was put <unk> #Number all tokens

The next step was create the FST Trasducer file called transducerTOK\_POS.txt, this file was create using the files TOK\_POS.probs

during this process the probabilities are changed by weights, using this formula :

weight(E) = -ln(P(E)),e =(w,t)

The the probability for unknown words is given by: P(<unk>| ti) =1/count ti



having the file transducerTOK\_POS.txt that is represents <initial step, final step, word, tag, cost> the next step is generate the FST. The WFST is create using the farcompilestrings, the result became the input to OpenGRM tool, the first command used is ngramcount that create the probability of a tag given the n-1 precedessors. in this paper was used 1-5 N-grams, the output is used to create the Language model, this paper use 5 smoothing methods, it allows create differents LM based on the combination between n-gram, smoothing method and cut-off

Test

There are some procedures to be executed in the test phase, initially was created three files, the files, tok\_actual.txt and sentences.txt will be used together with test output result to be evaluate...

instead the file sentences\_line.txt contains the sentences by line, for each line is used farcompilestrings that create a FAR file, it is splitted into a FST, each fst is compose with LM and Train FST. this process is made for each line and the result is salve into tok\_predicted.txt file

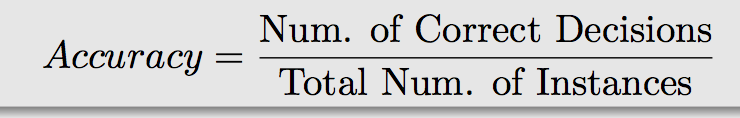
the files sentences.txt tok\_actual.txt tok\_predicted.txt are merged creating the triple is (word, actual-tag, predict-tag). the final step is evaluate it with conlleval.pl , it allows to produce the table with the quality of the models.

Results

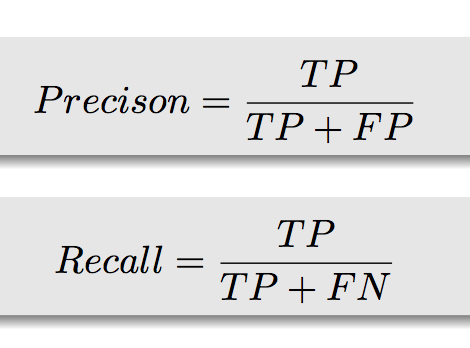
The idea is measure the output given by the SLU model

Evaluation Methods

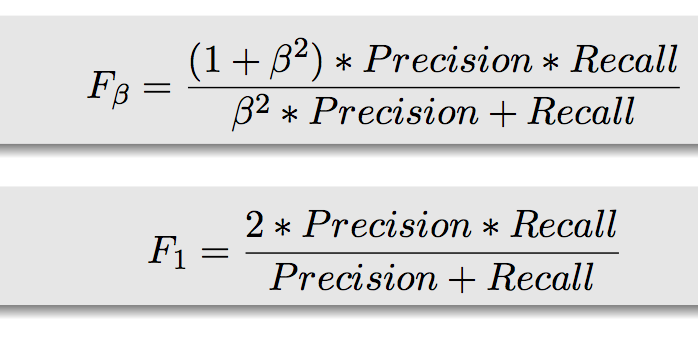
* **accuracy**: Accuracy refers to the closeness of a measured value to a standard or known value.



* **precision**: Precision refers to the closeness of two or more measurements to each other or also called positive predictive value, is the fraction of retrieved instances that are relevant



* **recall**: recall is the fraction of relevant instances that are retrieved in another way is a measure of how many truly relevant results are returned.
* **FB1**: is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score



Evaluation

IN the tabels above is possible see there evaluation, the tables was divided by smoothing method , the collon N-gram represent the ngram order considered, the next 3 colluns represent the evaluation methods accuracy, precision ,recall and F-mesuare.

enter the title, author name(s) and affiliation(s) across both columns (or, in the case of initial sub- mission, space for the names). Do not use footnotes for affiliations. Use the two-column format only when you begin the abstract.

**Title:** Place the title centered at the top of the first page, in a 15-point bold font. (For a complete guide to font sizes and styles, see Table 1.) Long titles should be typed on two lines without a blank line intervening. Approximately, put the title at 2.5 cm from the top of the page, followed by a blank line, then the author name(s), and the affiliation(s) on the following line. Do not use only initials for given names (middle initials are allowed). Do not format surnames in all capitals (*e.g.*, use “Mitchell,” not “MITCHELL”). Do not use format title and section headings in all capitals as well, except for proper names (such as “BLEU”) that are conventionally in all capitals. The affiliation should contain the author's complete address, and if possible, an electronic mail address. Start the body of the first page 7.5 cm from the top of the page.

The title, author names and addresses should be completely identical to those entered to the electronic paper submission website in order to maintain the consistency of author information among all publications of the conference. If they are different, the publication chairs may resolve the difference without consulting with you; so it is in your own interest to double-check that the information is consistent.

|  |  |
| --- | --- |
|  |  |

**Abstract:** Type the abstract at the beginning of the first column. The width of the abstract text should be smaller than the width of the columns for the text in the body of the paper by about 0.6cm on each side. Center the word **Abstract** in a 12 point bold font above the body of the abstract. The abstract should be a concise summary of the general thesis and conclusions of the paper. It should be no longer than 200 words. The abstract text should be in 10 point font.

**Text:** Begin typing the main body of the text immediately after the abstract, observing the two-column format as shown in the present document. Use 11 points for text. Indent when starting a new paragraph, about 0.16 in. This is accomplished with Right click, Styles, “ACL First Line”.

Sections

**Headings:** Type and label section and subsection headings in the style shown on the present document. Use numbered sections (Arabic numerals) in order to facilitate cross references. Number subsections with the section number and the subsection number separated by a dot, in Arabic numerals. In bold, use 11 points for subsection headings, 12 points for section headings. Do not number subsubsections.

**Citations:** Citations within the text appear in parentheses as ([Gusfield, 1997](#Gusfield1997)) or, if the author's name appears in the text itself, as Gusfield ([1997](#Gusfield1997)). Collapse multiple citations as in ([Gusfield, 1997](#Gusfield1997); [Aho and Ullman, 1972](#AhoUllman72)). Append lowercase letters to the year in cases of ambiguities. Treat double authors as in ([Aho and Ullman, 1972](#AhoUllman72)), but write as in ([Chandra et al., 1981](#ChandraEtAl1981)) when more than two authors are involved. Also refrain from using full citations as sentence constituents.

We suggest that instead of

“([Gusfield, 1997](#Gusfield1997)) showed that ...”

|  |  |
| --- | --- |
|  |  |

you use

“Gusfield ([1997](#Gusfield1997)) showed that ...”

**Hyperlinks:** Within-document and external hyperlinks are indicated with Dark Blue text, Color Hex #000099.In order to create hyperlinks between citations and references, as you insert each full reference in the References section, highlight it and then select Insert, Bookmark. Link back to the reference from its citations in the text by highlight the citation, right clicking, and selecting Insert, Cross-Reference, then selecting the Bookmark you’ve saved. Highlight the citation again to give make it dark blue (included in this theme), if it is not automatically applied. If there are problems saving the hyperlinks when you convert the document to PDF, use an online converter such as <http://go4convert.com>.

**Digital Object Identifiers:** As part of our work to make ACL materials more widely used and cited outside of our discipline, ACL has registered as a CrossRef member, as a registrant of Digital Object Identifiers (DOIs), the standard for registering permanent URNs for referencing scholarly materials. As of 2017, we are requiring all camera-ready references to contain the appropriate DOIs (or as a second resort, the hyperlinked ACL Anthology Identifier) to all cited works. Thus, please ensure that you use References that contain DOI or URLs for any of the ACL materials that you reference. Appropriate records should be found for most materials in the current ACL Anthology at <http://aclanthology.info/>.

As examples, we cite ([Goodman et al., 2016](#GoodmanEtAl2016)) to show you how papers with a DOI will appear in the bibliography. We cite ([Harper, 2014](#Harper2014)) to show how papers without a DOI but with an ACL Anthology Identifier will appear in the references.

As reviewing will be double-blind, the submit- ted version of the papers should not include the authors’ names and affiliations. Furthermore, self-references that reveal the author’s identity, *e.g.*,

“We previously showed ([Gusfield, 1997](#Gusfield1997)) ...”

should be avoided. Instead, use citations such as

“Gusfield ([1997](#Gusfield1997)) previously showed ... ”

**Please do not use anonymous citations** and donot include acknowledgements when submitting your papers. Papers that do not conform to theserequirements may be rejected without review.

**References:** Gather the full set of references together under the heading **References**; place the section before any Appendices. Arrange the references alphabetically by first author, rather than by order of occurrence in the text. Provide as complete a reference as possible, using a consistent format, such as the one for *Computational Linguistics* or the one in the *Publication Manual of the American Psychological Association* ([American Psychological Association, 1983](#APA83)). Use of full names for authors rather than initials is preferred. A list of abbreviations for common computer science journals can be found in the Association for Computing Machinery ([1983](#ACM83)) *Computing Reviews*.

The bibliography style described here roughly fits the American Psychological Association format, allowing regular citations, short citations and multiple citations as described above.

**Appendices:** Appendices, if any, directly follow the text and the references. Letter them in sequence and provide an informative title: **Appendix A. Title of Appendix**.

Footnotes

**Footnotes:** Put footnotes at the bottom of the page and use 9 point font. They may be numbered or referred to by asterisks or other symbols.[[1]](#footnote-2) Footnotes should be separated from the text by a line.[[2]](#footnote-3)

Figures and Tables

**Creating:** To create a new Figure or Table, insert a Text Box where you want it to appear (generally, centered at the top of a column close to where it is referred to) and then fill it in with the Figure (or Table). Highlight and right click to add Caption, with the ACL Caption style, which places 10 pt below and above the caption.

**Placement:** Place figures and tables in the paper near where they are first discussed, as close as possible to the top of their respective column. Wide figures and tables may run across both columns and should be placed at the top of a page.

|  |
| --- |
|  |
| Figure 1: Figure caption**.** |

In MSWord, authors can place a Figure (*e.g.*, a graphic and its caption) inside the rows of a 2 x 1 table (2 rows and 1 column) with invisible borders. Specify table positioning by right-clicking its handle in the upper left corner. Place the image in the center of the first row, and the caption in the center of the second row.

**Captions:** Provide a caption for every table and figure; number each one sequentially in the form: “Figure 1: Figure caption.”, “Table 1: Table caption.” Type the captions of the figures and tables below the body, using 11 point text.

**Numbering:** To update numbering, highlight all the relevant text (*e.g.*, **Ctrl-A + F9**). This will update all the numbering applicable to tables, figures, equations, and headings.

**Cross-referencing:** To add a cross reference to a figure or table:

* Place the mouse pointer at the location where you wish to add the cross-reference.
* Click on the **Insert** menu, (then click **Reference**), and then **Cross-reference** in the **Links** panel.
* In the **Cross-reference** dialog box, click the caption to which you are building the text reference.
* For a figure, under **Reference Type**, click **Figure**.
* Under Insert Reference To, click Only Label and Number, then click OK.
* Once the reference is in place, apply the ‘Normal’ font style (size 11, no bold face).
* This is an example reference to Figure 1.

Equations

An example equation is shown below:

(1)

To add new equations, authors are encouraged to copy this existing equation line, and then replace with the new equation. The numbering and alignment of equation line elements is automatic. To update equation numbering, press **Ctrl-A + F9**. Note: this will only update the number to the right of the equation; to update numbering within the text you must create a cross-reference.

|  |  |
| --- | --- |
|  |  |

**Cross-referencing:** To create a cross-reference for an equation:

* Create a bookmark for it.
* Select the number to the right of the equation. Go to **Insert**, **Bookmark** (in the **Links** panel),andthen create a name for your equation. Press **Add** to create the bookmark.
* To refer back, place the mouse pointer at the location where you wish to add the cross reference.
* Go to **Insert, Cross-reference** (in the **Links** panel).In the dialogue box, select **Bookmark** and **Bookmark Text** from each dropdown list. Uncheck **Insert as Hyperlink**, then click **OK**.
* This will make it such that whenever a new equation is added, the references to the equation will be updated when **Ctrl-A + F9** is pressed.
* This an example cross-reference to Equation (1).

Accessibility

In an effort to accommodate the color-blind (as well as those printing to paper), grayscale readability for all accepted papers is encouraged. Color is not forbidden, but authors should ensure that tables and figures do not rely solely on color to convey critical distinctions. A simple criterion: All curves and points in your figures should be clearly distinguishable without color.

From Submitted to Camera-Ready

To remove submission formatting for the camera-ready document, delete the header and footer on the first page (double clicking in the area, then delete the text). This will remove them from the rest of the document. To remove the ruler, highlight the lists of numbers and delete.

Translation of non-English Terms

It is also advised to supplement non-English characters and terms with appropriate transliterations and/or translations since not all readers understand all such characters and terms. Inline transliteration or translation can be represented in the order of:

original-form

transliteration

“translation”.

Length of Submission

The ACL 2017 main conference accepts submissions of long papers and short papers. Long papers may consist of up to eight (8) pages of content, plus unlimited pages for references. Upon acceptance, final versions of long papers will be given one additional page – up to nine (9) pages with unlimited pages for references – so that reviewers’ comments can be taken into account. Short papers may consist of up to four (4) pages of content, plus unlimited pages for references. Upon acceptance, short papers will be given five (5) pages in the proceedings and unlimited pages for references.

|  |  |
| --- | --- |
|  |  |

For both long and short papers, all figures and tables that are part of the main text must be accommodated within these page limits, observing the formatting instructions given in the present document. Supplementary material in the form of appendices does not count towards the page limit.

However, note that supplementary material should be supplementary (rather than central) to the paper, and that reviewers may ignore supplementary material when reviewing the paper (see Appendix A). Papers that do not conform to the specified length and formatting requirements are subject to be rejected without review.

Workshop chairs may have different rules for allowed length and whether supplemental material is welcome. As always, the corresponding call for papers is the authoritative source.

STREAM Tools

This Microsoft Word file was updated in 2016 with STREAM Tools, designed for creating well-formatted reports and papers with Microsoft Word (Mamishev, 2010; Mamishev, 2013).

Acknowledgments

The acknowledgements should go immediately before the references. Do not number the acknowledgments section. Do not include this section when submitting your paper for review.

References

Alfred. V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling, volume 1*. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual.* American Psychological Association, Washington, DC.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. [Alternation](http://dl.acm.org/citation.cfm?doid=322234.322243). *Journal of the Association for Computing Machinery*, 28(1):114-133. https://doi.org/10.1145/322234.32224.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503-512.

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. [Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing](http://aclweb.org/anthology/P16-1001). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1–11. <https://doi.org/10.18653/v1/P16-1001>.

|  |  |
| --- | --- |
|  |  |

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Mary Harper. 2014. [Learning from 26 languages: Pro- gram management and science in the babel program](http://aclweb.org/anthology/C14-1001). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, page 1. <http://aclweb.org/anthology/C14-1001>.

Alexander V. Mamishev and Murray Sargent. 2013. *Creating Research and Scientific Documents Using Microsoft Word*. Microsoft Press, Redmond, WA.

Alexander V. Mamishev and Sean D. Williams. 2010. *Technical Writing for Teams: The STREAM Tools Handbook*. Wiley-IEEE Press, Hoboken, NJ.

1. Supplementary Material

ACL 2017 also encourages the submission of supplementary material to report preprocessing decisions, model parameters, and other details necessary for the replication of the experiments reported in the paper. Seemingly small preprocessing decisions can sometimes make a large difference in performance, so it is crucial to record such decisions to precisely characterize state-of-the-art methods.

Nonetheless, supplementary material should be supplementary (rather than central) to the paper. **Submissions that misuse the supplementary material may be rejected without review.** Essentially, supplementary material may include explanations or details of proofs or derivations that do not fit into the paper, lists of features or feature templates, sample inputs and outputs for a system, pseudo-code or source code, and data. (Source code and data should be separate uploads, rather than part of the paper).

The paper should not rely on the supplementary material: while the paper may refer to and cite the supplementary material and the supplementary material will be available to the reviewers, they will not be asked to review the supplementary material.

Appendices (*i.e.*, supplementary material in the form of proofs, tables, or pseudo-code) should come after the references, as shown here.

1. **Multiple Appendices**

…can be gotten by using more than one section. We hope you won't need that.

1. This is how a footnote should appear. [↑](#footnote-ref-2)
2. Note the line separating the footnotes from the text. [↑](#footnote-ref-3)