Spelling Error Annotation Guidelines For Consumer Health Text

This document describes the specific types of information that are annotated for the spelling error detection and correction task on consumer health text. Examples found in the consumer health text are provided. This document is organized into three sections:

1. Introduction
2. Annotation
3. Quality assurance

1. Introduction

The purposes of the spelling errors annotation task is to establish gold standard for evaluating spelling error detection and correction. The collection consists of 224 consumer health questions with high count of OOV (out of vocabulary) tokens. The OOV spelling errors in the questions are pre-annotated. The brat rapid annotation tool is used for manual annotation.

2. Annotation

Annotate each spelling error with a corresponding error type. When needed, provide the correct spelling in the NOTE field of the annotation tool. For example, annotate "knwo", in document 1-131373925 as Misspelling and enter the corrected spelling "know" in the NOTE field of the annotation. The corrected text is required in NOTES for all tags except ToSplitOnPunct, ToMerge, WordExists, Unknown.and Garbage.

2.1 Spelling error tag types

- OutOfVocabulary:
  - The words that could not be found in the SPECIALIST Lexicon are automatically pre-annotated with an OutOfVocabulary tag. Note that some spelling errors might not be pre-annotated with OOV; please carefully look for those and annotate as needed.
  - All OutOfVocabulary TAGs will be re-assigned to new TAG(s) when the annotation task is completed.

- WordExists:
  - Use this TAG if the source text is a valid word, but is initially marked as "OutOfVocabulary" TAG. For example, "Tecfidera", in document 1-118268098, is initially tagged as OutOfVocabulary. Change the tag to WordExists.
  - **Do not** add corrected text to NOTES.
  - **Add** only comments/notes to NOTES.
  - Multiple TAGs are **not** allowed.
  - Examples:

ID	**Source Text**	**Corrected Text (Do not add to NOTES)**	**Document**

WE1	Tecfidera	Tecfidera	1-118268098
WE2	Skype	Skype	1-130905055
WE3	fildena	fildena	1-132540057

- ToSplit:
  - Use this tag if space(s) need to be added to the source text for correction, except for the case of ToSplitOnPunct (see below).
  - Also use this tag if punctuation needs to be replaced with a space (SP11, SP12 and SP13).
  - Add corrected text to NOTES.
  - Include multiple splits (SP4, SP5 and SP8) in one tag.
  - Multiple TAGs are allowed.
  - Examples:

ID	Source Text	Corrected text	Document
SP1	10years	10 years	1-132156225
SP2	every12	every 12	1-132464205
SP3	hotflashes	hot flashes	1-120029095
SP4	brokenribscantsleepatnight	broken ribs cant sleep at night	1-131053995
SP5	Amlodipine5mgs	Amlodipine 5 mgs	1-135402005
SP6	1.2mm	1.2 mm	1-118342705
SP7	40.00for	40.00 for	1-132464205
SP8	.5&75mg	.5 & 75 mg	1-131195919
SP9	Herpes(_)	Herpes (_)	1-132123145
SP10	clot?...is	clot?... is	1-134591345
SP11	neck-lesion	neck lesion	1-120050175
SP12	celiac.disease	celiac disease	1-123259335

Use ToSplit (not ToSplitOnPunct) for examples SP6-SP10 in the above table (even though they include punctuation) because:

- split is combined with non-punctuation split (SP6-SP8)
- no space is added after some punctuation (SP6-SP10)
- a space is added before the punctuation (SP8-SP9)

- ToSplitOnPunct:
  - This TAG is a special case of ToSplit. Use this TAG only for splits after all punctuation in the source text. Annotate the punctuation and the following word (see bold text in examples).
  - **Do not** add corrected text to NOTES.
  - Computer programs generate the corrected text automatically in post process.
  - Include multiple splits after the punctuation (Example, SPP3)

- Multiple TAGs are allowed.
- Examples:

ID	Source Text	Corrected text (Do not add to NOTES)	Document
SPP1	drugs**.What**	drugs. What	1-135402005
SPP2	CONDITION**.SHE**	CONDITION. SHE	1-123227295
SPP3	FEVER**,PAIN.NOT**	FEVER, PAIN. NOT	1-123227295

- ToMerge:
  - Use this TAG if all space(s) or punctuation need to be deleted from the source text for correction.
  - **Do not** add corrected text to NOTES.
  - Computer programs generate the corrected text automatically in post process
  - Include multiple merges (Example: MG2)
  - Multiple TAGs are allowed.
  - Examples:

ID	Source Text	Corrected Text (Do not add to NOTES)	Document
MG1	POLY UREA	POLYUREA	1-123227295
MG2	stiff n ess	stiffness	1-119980475

- RealWord:
  - Use this TAG if the source text is a valid word but the meaning of the source text and corrected text are different.
  - **Add** corrected text to NOTES.
  - Multiple TAGs are allowed.

Examples:

ID	Source Text	Corrected Text	Document
RW1	POLYUREA	polyuria	1-123227295
RW2	leaver	liver	1-118259395
RW3	lumber	lumbar	1-121798454
RW4	their	there	1-123706915
RW5	its	it is	1-123737547
RW6	cant	cannot	1-137191537
RW7	what ever	whatever	1-132540057

- Informal:
  - Use this TAG if the source text is an informal expression. Informal expressions include contractions, shorthand, abbreviations, acronyms, etc. A contraction with a missing apostrophe is treated as informal (not a misspelling) EXCEPT when it is a real word (e.g., cant). Mark can't as informal, but mark cant as RealWord and expand both to cannot. Very common units of measure (eg., mg, yrs) do not need to be annotated.
  - For acronyms and abbreviations, precede the expansion in the NOTES with "ACR:" or "ABB:".
  - Add corrected text to NOTES.
  - Multiple TAGs are allowed.

    Examples:

ID	Source Text	Corrected Text	Document
IF1	didnt	did not	1-120035472
IF2	pls	please	1-118315685
IF3	u	you	1-120035472
IF4	COORDIN	ABB: coordination	1-120014465
IF5	GI	ACR: gastrointestinal	1-132340905
IF6	I've	I have	1-119980475

- Misspelling:
  - Use this TAG if the source text is misspelled (not real words or informal expression)
  - Add corrected text to NOTES.
  - Multiple TAGs are allowed.
  - Examples:

ID	Source Text	Corrected Text	Document
MS1	presure	pressure	1-118259395
MS2	0,71	0.71	1-120035055

- Punctuation:
  - Use this TAG only for syntax errors on punctuation. Correct punctuation as follows:
    - Sentences should end with a period (.) or question mark (?). Do not add an exclamation mark (!).
    - Add/modify punctuation if obviously missing or used incorrectly.
    - Parenthesis, square brackets, curly braces should be symmetric. Add them if they are obviously missing.
  - Add corrected text to NOTES.

  - Multiple TAGs are allowed.
  - Examples:

ID	Source Text	Corrected text	Document
PT1	it	it.	1-133501085
PT2	failure	failure?	1-133501085
PT4	(3 - 500mg tab,	tab),	1-121789105

The grayed out text in PT4 is used to illustrate the missing right parenthesis.

- Unknown:
  - Use this TAG for cases where the correct text is unknown.
  - Guesses could be added to NOTES.
  - Multiple TAGs are allowed.

Example:

ID	Source Text	Suggestion/Guess text	Document
UK1	Cronin		1-135400012
UK2	wit	this looks like a fragment of without	1-118274802

We do not know what the correct text is in UK1, thus no guess text is added to NOTES.

- Garbage:
  - Use this TAG for cases where the source text is not real word but no need to be corrected, such as URL, telephone numbers, etc.
  - **Do not** add corrected text to NOTES.
  - Multiple TAGS are **not** allowed.

Example:

ID	Source Text	Corrected Text (Do not add to NOTES)	Document
GB1	0115-0672-50		1-136162032
GB2	na…na…		1-118274802

2.1.1 Multiple annotations on a span of text

TAGs are categorized into four groups according to the order in the annotation process. The corrected text of the 1st TAG is used as the source text for the 2nd TAG, and so on. Each split token has its own TAG after ToSplit or ToSplitOnPunct. The table below summarizes the requirements of NOTES and multiple TAGs priority group for each TAG.

TAG	NOTES	Multiple TAGs Priority Group
ToSplit	Add corrected text	1st
ToSplitOnPunct	**None**	1st
ToMerge	**None**	1st
RealWord	Add corrected text	2nd
Informal	Add corrected text	2nd
Misspelling	Add corrected text	2nd
Punctuation	Add corrected text	3rd
Unknown	**None** or add suggestion/guess text	3rd
OutOfVocabulary	**None**	No
WordExists	**None** or add comments	No
Garbage	**None**	No

The rules for multiple annotations on the same token are described below:

- **1st priority:**
    - Includes three TAGs: ToSplit, ToSplitOnPunct and ToMerge that allow only deleting or adding space(s).

- **2nd priority group:**
    - Includes three TAGs: RealWord, Informal and Misspelling
    - **Only 1 TAG** of RealWord, Informal and Misspelling should be marked in this 2nd priority group. For example, use only 1 TAG, Misspelling for source text, “dn't”, and add correct text, “do not”, to the NOTES in document 1-122992595 (the source text is an invalid word).
    - Use the following criteria on the source text to decide which TAG should be marked in this group:
        - If the source text is a valid word, mark RealWord
        - If the source text is not a valid word, but it is commonly used as informal expression, mark Informal

    Example:

ID	Source Text	Correct text	TAG	Document
AT1	cant	cannot	RealWord	1-131053995
AT2	dont	do not	Informal	1-123940155
AT3	can,t	cannot	Misspelling	1-134864077

- **3rd priority group**:
    - Includes two TAGs: Punctuation and Unknown
    - Use these TAGs after the 1st or 2nd priority group in multiple TAGs.

- **No priority group**:
  - Includes three TAGs: WordExists, OutOfVocabulary and Garbage.
  - These three TAGs are not allowed in multiple TAGs.

The following table shows examples of multiple tags:

ID	1st Tag Source Text	1st TAG & NOTES	2nd Tag Source Text	2nd TAG & NOTES	Document
MT1	leaver&high	ToSplit leaver & high	leaver	RealWord liver	1-118259395
MT2	brokenribscantsleepatnight	ToSplit broken ribs cant sleep at night	cant	RealWord cannot	1-131053995
MT3	work.plz	ToSplitOnPunct (work. plz)*	plz	Informal please	1-131991125
MT4	body-What	ToSplitOnPunct (body- What)*	body-	Punctuation body.	16523
MT5	MEDACATION,THES	ToSplitOnPunct (MEDACATION, THES)*	MEDACATION	Misspelling medication	1-131093935
			THES	Misspelling These	
MT6	POLY UREA	ToMerge (POLYUREA)*	POLYUREA	RealWord polyuria	1-123227295
MT7	bumb	Misspelling bump	bump	Punctuation bump.	1-132842405

*There is no need to add correct text for TAGs of ToSplitOnPunct and ToMerge for examples MT3-MT6. However, the correct text is shown within the parenthesis under the column of the 1st TAG & NOTES for illustration purpose.

III. Quality assurance

The following steps ensure the quality of the annotation. The result is then used to generate the gold standard for the test set.

- The annotators reconcile their annotation differences to produce the final annotation.
- Make sure there are no OutOfVocabulary TAGs in the reconciled annotations.
- Manually review all Unknown and Garbage TAGs.