“NLP Evaluation”
Lecture review

Martin Rajman

Laboratoire d’Intelligence Artificielle

Faculté I&C

The NLP evaluation protocol

1. Define a control task
2. Produce a reference
3. Assess the quality of the reference
4. Use the reference to evaluate NLP systems
5. Discuss and publish the results

Control task

If possible, model the control task as a N-ary classification

…Why?...

Control task: the N-ary classification case

If the (control) task is modeled as a N-ary classification, the outputs produced by *any pair of annotators* (human expert or automated system) can be synthesized in a form of a confusion matrix, on which all the evaluation can be performed

For example, for 2 systems performing a ternary classification

		system1		
		class1	**class2**	**class2**
system2	**class1**	*n11*	*n12*	*n13*
	class2	*n21*	*n22*	*n23*
	class3	*n31*	*n32*	*n33*

where nij represents the number of cases, annotated by class i by system 2, and by class j by system 1

Control task: the N-ary classification case (2)

In particular, if the pair of annotators (annotating the inputs present in the selected reference) consists of:

- two human experts, the resulting confusion matrix will be used to compute the "Inter Annotator Agreement" (IAA) that measures the quality of the reference
- a human expert and an automated system, the resulting confusion matrix will be used to compute various evaluation metrics that measure the performance of the automated system

Control task: Extractive summarization

"Extractive summarization" relies on the following 3 steps:

1. The text to summarize is first decomposed into sentences;
2. For each of the sentences present in the original text, a binary "keep / not keep" decision is taken, with the constraint that only (max) Ns "keep" decisions can be taken;
3. The resulting "raw" summary is potentially reformulated to increase readability.

Example

Consider the following text:

Germany's left-leaning Social Democratic Party (SPD) has won the most seats in the country's federal election, preliminary results show, but it will be some time before the makeup of the new government is known. The SPD claimed a narrow victory over the Christian Democratic Union (CDU), the centrist-right, conservative party of outgoing Chancellor Angela Merkel, according to the "Federal Returning Officer" responsible for overseeing Federal elections. The Federal Returning Officer website said the SPD won 25.7% of the vote, followed by the CDU/CSU bloc which garnered 24.1%, and the Green Party with 14.8% of votes, after a count of all 299 of Germany's "constituencies" or electoral districts.

The election ends Merkel's 16-year stint in the top job, but her successor won't be decided until a coalition deal is negotiated. The SPD will now begin negotiations to form the new government, a process that could take weeks -- or even months. After Merkel's election win in September 2017, it took more than five months for a government to be formed.

Though the preliminary count gives the SPD a small lead over its closest rivals, the results mark a significant improvement for the party that took 20.5% of the vote in the last election in 2017. As votes were counted, party leader Olaf Scholz called the outcome a "great success.“ The 63-year-old politician has served as the vice-chancellor and German finance minister in Merkel's grand coalition government since 2018, earning him increased visibility as he navigated Germany's economic response to the pandemic. Loud applause and cheering from jubilant party supporters interrupted him as he spoke.

Step 1: Splitting into sentences

S01: Germany's left-leaning Social Democratic Party (SPD) has won the most seats in the country's federal election, preliminary results show, but it will be some time before the makeup of the new government is known.

S02: The SPD claimed a narrow victory over the Christian Democratic Union (CDU), the centrist-right, conservative party of outgoing Chancellor Angela Merkel, according to the "Federal Returning Officer" responsible for overseeing Federal elections.

S03: The Federal Returning Officer website said the SPD won 25.7% of the vote, followed by the CDU/CSU bloc which garnered 24.1%, and the Green Party with 14.8% of votes, after a count of all 299 of Germany's "constituencies" or electoral districts.

S04: The election ends Merkel's 16-year stint in the top job, but her successor won't be decided until a coalition deal is negotiated.

S05: The SPD will now begin negotiations to form the new government, a process that could take weeks -- or even months.

S06: After Merkel's election win in September 2017, it took more than five months for a government to be formed.

S07: Though the preliminary count gives the SPD a small lead over its closest rivals, the results mark a significant improvement for the party that took 20.5% of the vote in the last election in 2017.

S08: As votes were counted, party leader Olaf Scholz called the outcome a "great success.“

S09: The 63-year-old politician has served as the vice-chancellor and German finance minister in Merkel's grand coalition government since 2018, earning him increased visibility as he navigated Germany's economic response to the pandemic.

S10: Loud applause and cheering from jubilant party supporters interrupted him as he spoke.

Step 2: Annotating the sentences

[X]S01: Germany's left-leaning Social Democratic Party (SPD) has won the most seats in the country's federal election, preliminary results show, but it will be some time before the makeup of the new government is known.

[ ]S02: The SPD claimed a narrow victory over the Christian Democratic Union (CDU), the centrist-right, conservative party of outgoing Chancellor Angela Merkel, according to the "Federal Returning Officer" responsible for overseeing Federal elections.

[X]S03: The Federal Returning Officer website said the SPD won 25.7% of the vote, followed by the CDU/CSU bloc which garnered 24.1%, and the Green Party with 14.8% of votes, after a count of all 299 of Germany's "constituencies" or electoral districts.

[ ]S04: The election ends Merkel's 16-year stint in the top job, but her successor won't be decided until a coalition deal is negotiated.

[ ]S05: The SPD will now begin negotiations to form the new government, a process that could take weeks -- or even months.

[ ]S06: After Merkel's election win in September 2017, it took more than five months for a government to be formed.

[X]S07: Though the preliminary count gives the SPD a small lead over its closest rivals, the results mark a significant improvement for the party that took 20.5% of the vote in the last election in 2017.

[ ]S08: As votes were counted, party leader Olaf Scholz called the outcome a "great success.“

[ ]S09: The 63-year-old politician has served as the vice-chancellor and German finance minister in Merkel's grand coalition government since 2018, earning him increased visibility as he navigated Germany's economic response to the pandemic.

[ ]S10: Loud applause and cheering from jubilant party supporters interrupted him as he spoke.

…Select the sentences you believe should be kept... (how to do this?)

Step 2: Annotating the sentences (2)

[X]S01: Germany's left-leaning Social Democratic Party (SPD) has won the most seats in the country's federal election, preliminary results show, but it will be some time before the makeup of the new government is known.

[ ]S02: The SPD claimed a narrow victory over the Christian Democratic Union (CDU), the centrist-right, conservative party of outgoing Chancellor Angela Merkel, according to the "Federal Returning Officer" responsible for overseeing Federal elections.

[X]S03: The Federal Returning Officer website said the SPD won 25.7% of the vote, followed by the CDU/CSU bloc which garnered 24.1%, and the Green Party with 14.8% of votes, after a count of all 299 of Germany's "constituencies" or electoral districts.

[ ]S04: The election ends Merkel's 16-year stint in the top job, but her successor won't be decided until a coalition deal is negotiated.

[ ]S05: The SPD will now begin negotiations to form the new government, a process that could take weeks -- or even months.

[ ]S06: After Merkel's election win in September 2017, it took more than five months for a government to be formed.

[ ]S07: Though the preliminary count gives the SPD a small lead over its closest rivals, the results mark a significant improvement for the party that took 20.5% of the vote in the last election in 2017.

[ ]S08: As votes were counted, party leader Olaf Scholz called the outcome a "great success.“

[ ]S09: The 63-year-old politician has served as the vice-chancellor and German finance minister in Merkel's grand coalition government since 2018, earning him increased visibility as he navigated Germany's economic response to the pandemic.

[ ]S10: Loud applause and cheering from jubilant party supporters interrupted him as he spoke.

…correspond to my personal choices...

Step 3: Reformulating the raw summary

Germany's left-leaning Social Democratic Party (SPD) has won the most seats in the country's federal election, preliminary results show, but it will be some time before the makeup of the new government is known.

The Federal Returning Officer (who's s/he?) website said the SPD won 25.7% of the vote, followed by the CDU/CSU (what's this?) bloc which garnered 24.1%, and the Green Party with 14.8% of votes, after a count of all 299 of Germany's "constituencies" or electoral districts.

➔

Germany's left-leaning Social Democratic Party (SPD) has won the most seats in the country's federal election, preliminary results show, but it will be some time before the makeup of the new government is known.

The website of the Federal Returning Officer responsible for overseeing Federal elections said the SPD won 25.7% of the vote, followed by the Christian Democratic Union (CDU) bloc which garnered 24.1%, and the Green Party with 14.8% of votes, after a count of all 299 of Germany's "constituencies" or electoral districts.

Notes: Splitting into sentences (2)

The extractive summarization may operate on clauses instead of sentences:

C01: Germany's left-leaning Social Democratic Party (SPD) has won the most seats in the country's federal election

C02: (…) preliminary results show

C03: (…) but it will be some time before the makeup of the new government is known.

C04: The SPD claimed a narrow victory over the Christian Democratic Union (CDU), the centrist-right, conservative party of outgoing Chancellor Angela Merkel

C05: (…) according to the "Federal Returning Officer" responsible for overseeing Federal elections.

C06: The Federal Returning Officer website said (…)

C07: the SPD won 25.7% of the vote after a count of all 299 of Germany's "constituencies" or electoral districts.

C08: (…) followed by the CDU/CSU bloc and the Green Party with 14.8% of votes

C09: (…) which garnered 24.1%,

C10: The election ends Merkel's 16-year stint in the top job

C11: (…) but her successor won't be decided

C12: (…) until a coalition deal is negotiated.

C13: The SPD will now begin negotiations to form the new government

C14: (…) a process that could take weeks -- or even months.

C15: After Merkel's election win in September 2017

C16: (…) it took more than five months

C17: (…) for a government to be formed.

C18: Though the preliminary count gives the SPD a small lead over its closest rivals

C19: (…) the results mark a significant improvement for the party that took 20.5% of the vote in the last election in 2017.

C20: As votes were counted,

C21: (…) party leader Olaf Scholz called the outcome a "great success."

C22: The 63-year-old politician has served as the vice-chancellor and German finance minister in Merkel's grand coalition government since 2018

C23: (…) earning him increased visibility as he navigated Germany's economic response to the pandemic.

C24: Loud applause and cheering from jubilant party supporters interrupted him

C25: (…) as he spoke.

… But this will make the reformulation step No3 substantially harder (why?)

Note: How to control the length of the summary?

What are your suggestions?

The NLP evaluation protocol

1. Define a control task
2. Produce a reference
3. Assess the quality of the reference
4. Use the reference to evaluate NLP systems
5. Discuss and publish the results

Producing a reference

Characteristics of a good reference?

Producing a reference

Characteristics of a good reference?

→ Relevant enough (why?)
→ Large enough (why?)
→ Consensual enough (what is this? why?)

Example: Extractive summarization

- We already have two expert human annotations…

Sentence	Annotator 1	Annotator 2
1	keep	keep
2	not keep	not keep
3	keep	keep
4	not keep	not keep
5	not keep	not keep
6	not keep	not keep
7	not keep	keep
8	not keep	not keep
9	not keep	not keep
10	not keep	not keep

Note: Why can both be considered as expert annotations?

The NLP evaluation protocol

1. Define a control task
2. Produce a reference
3. Assess the quality of the reference
4. Use the reference to evaluate NLP systems
5. Discuss and publish the results

Example: Using the Kappa as IAA measure

- The confusion matrix to consider is:

		annotator1	
		keep	**not keep**
annotator2	**keep**		
	not keep		

raw agreement =
Proba (1) says “keep” =
Proba (1) says “not keep” =
Proba (2) says “keep” =
Proba (2) says “not keep” =
Proba (1)&(2) agree on “keep” =
Proba (1)&(2) agree on “not keep” =
by-chance agreement =

$$\text{Kappa} = \frac{\text{(raw agreement – by-chance agreement)}}{\text{(1 – by-chance agreement)}} =$$

- Conclusion?

Example: Using the Kappa as IAA measure

- The confusion matrix to consider is:

		annotator1	
		keep	**not keep**
annotator2	**keep**	2	1
	not keep	0	7

raw agreement = 90%
Proba (1) says “keep” = 20%
Proba (1) says “not keep” = 80%
Proba (2) says “keep” = 30%
Proba (2) says “not keep” = 70%
Proba (1)&(2) agree on “keep” = 6%
Proba (1)&(2) agree on “not keep” = 56%
by-chance agreement = 62%

$$\text{Kappa} = \frac{\text{(raw agreement – by-chance agreement)}}{\text{(1 – by-chance agreement)}} = 0.74$$

- Conclusion?

The NLP evaluation protocol

1. Define a control task
2. Produce a reference
3. Assess the quality of the reference
4. Use the reference to evaluate NLP systems
5. Discuss and publish the results

Example: Evaluating a system

If we use “annotator 1” as “system1” and “annotator 2” as “reference”

- The corresponding annotations are:

Sentence	System1	Reference
1	keep	keep
2	not keep	not keep
3	keep	keep
4	not keep	not keep
5	not keep	not keep
6	not keep	not keep
7	not keep	keep
8	not keep	not keep
9	not keep	not keep
10	not keep	not keep

Example: Evaluating a system (2)

- … And the corresponding confusion matrix is:

		reference	
		keep	**not keep**
system1	**keep**	2	0
	not keep	1	7

- which allows to evaluate “system1” with various metrics derived from the confusion matrix:

  accuracy = 90%

  Precision = 100%

  Recall = 67%

  F-score = 80%

Which metric should be used? Accuracy? Precision/Recall? F-score? Why?

The NLP evaluation protocol

1. Define a control task
2. Produce a reference
3. Assess the quality of the reference
4. Use the reference to evaluate NLP systems
5. Discuss and publish the results

Example: Discussing the statistical significance

Assume that:

- We are evaluating “system1” based on the F-score, and
- We obtain the following F-score values through a 2x5 cross-validation:

F-score	1	2	3	4	5	6	7	8	9	10
system1	67%	63%	68%	65%	69%	61%	64%	67%	68%	65%

What information about the “true” F-score of “system1” can we derive from the available F-score values?

Example: Discussing the statistical significance (2)

Compute the associated confidence interval. To do so:

1. Compute the empirical mean $\overline{X}$

$$\overline{X} = \frac{1}{N} \sum_i X_i$$

where N is the number of available F-score values $X_1, \ldots, X_N$

2. Compute the empirical unbiased standard deviation s

$$s = \sqrt{\frac{1}{N-1} \sum_i (X_i - \overline{X})^2}$$

3. The formula for the c%-confidence interval is then

$$\left[\overline{X} - t_{N,c} \frac{s}{\sqrt{N}}, \overline{X} + t_{N,c} \frac{s}{\sqrt{N}}\right], \text{where } t_{N,c} \text{is a tabulated constant}$$

t(N,c) : (1-c)/2 level two-sided t-value of a Student t-distribution with N-1 degrees of freedom

c \ N	10	20	30	>>30
95%	2.262	2.093	2.045	1.960
99%	3.250	2.861	2.756	2.576

Example: Discussing the statistical significance (3)

F-score	1	2	3	4	5	6	7	8	9	10	Mean	Stdev
system1	67%	63%	68%	65%	69%	61%	64%	67%	68%	65%	65.70%	0.025

$\overline{X}$ = 65.70%

s = 0.025

The 95%-confidence interval for the true F-score is:

[63.88%, 67.52%]

Example: Discussing the statistical significance (4)

Assume that:

- We are evaluating another system ("system2") based on the F-score, and
- We obtain the following F-score values through the same 2x5 cross-validation:

F-score	1	2	3	4	5	6	7	8	9	10	Mean
system1	67%	63%	68%	65%	69%	61%	64%	67%	68%	65%	65.70%
system2	65%	66%	67%	68%	61%	62%	63%	66%	69%	60%	64.70%

We may hypothesize that "system1" is better than "system2", but what information about the comparison of the two systems can we derive from the available F-score values?

Example: Discussing the statistical significance (5)

Perform a statistical test on the F-score difference. To do so:

1. Compute the empirical mean difference $\overline{D}$

$$\overline{D} = \frac{1}{N} \sum_i D_i \text{ with } D_i = X_i^{(1)} - X_i^{(2)}$$

where $X^{(1)}_1$, …, $X^{(1)}_N$ (resp. $X^{(2)}_1$, …, $X^{(2)}_N$) are the N F-score values available for "system" (resp. "system2")

2. Compute the corresponding empirical unbiased standard deviation s

$$s = \sqrt{\frac{1}{N-1} \sum_i (D_i - \overline{D})^2}$$

T(N,c) : (1-c) level one-sided t-value of a Student t-distribution with N-1 degrees of freedom

c \ N	10	20	30	>>30
95%	1.833	1.729	1.699	1.645
99%	2.821	2.539	2.462	2.326

3. Compute the statistic of the one-sided paired Student t-test $T = \frac{\overline{D}\sqrt{N}}{s}$ and compare it to the threshold value $T_{N,c}$ tabulated for the test

Example: Discussing the statistical significance (6)

F-score	1	2	3	4	5	6	7	8	9	10	Mean	Stdev
system1	67%	63%	68%	65%	69%	61%	64%	67%	68%	65%	65.70%	0.025
system2	65%	66%	67%	68%	61%	62%	63%	66%	69%	60%	64.70%	0.031
(1)-(2)	2%	-3%	1%	-3%	8%	-1%	1%	1%	-1%	5%	1.00%	0.034

$\overline{D}$ = 1%

s = 0.034

T = 0.921

$T_{10,95\%}$ = 1.833 thus T < $T_{10,95\%}$

➔ at a 95%-confidence level, we cannot conclude that “system1” is indeed better than “system2”…