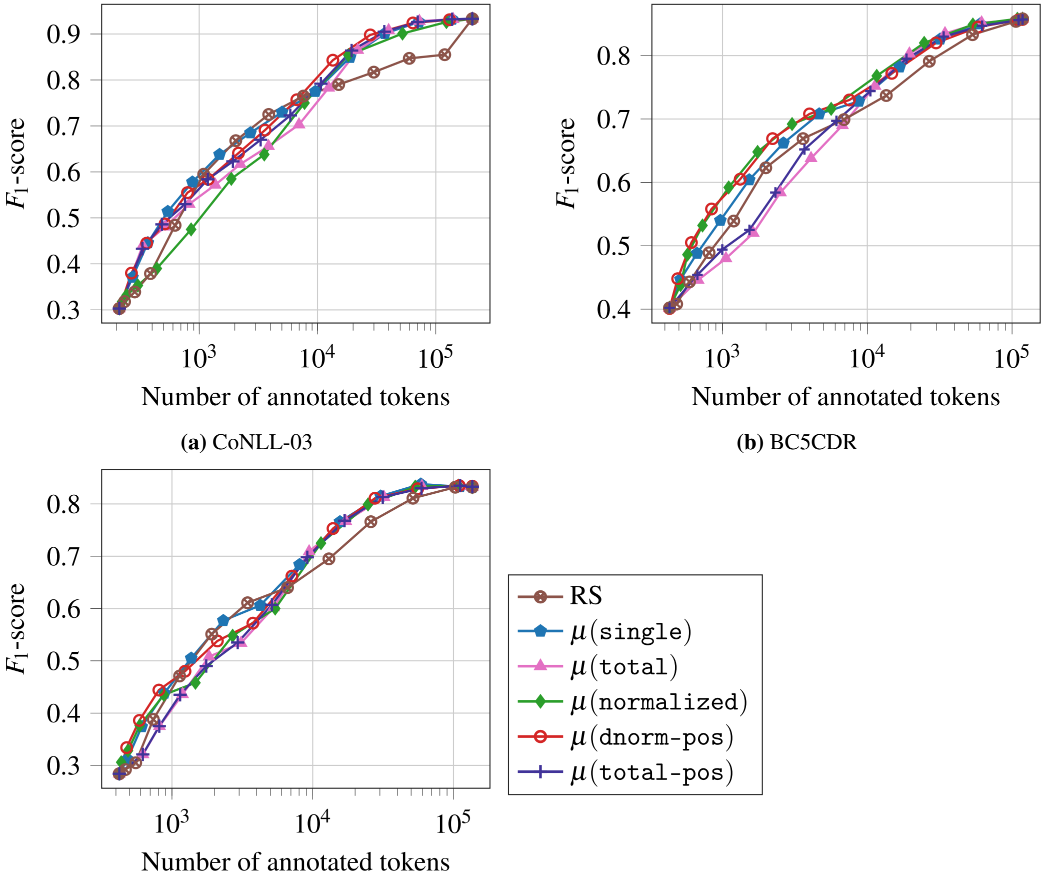
Appendix S1.

Table S1. The F_1-scores of the baseline methods (first three rows) and uncertainty scores together with their corresponding sentence score aggregation strategies average F_1-scores in the last four iterations before convergence. The percentage of sentences queried labeled set is provided for corresponding iterations under the iteration number. The reported deviations are the corresponding *standard error of the mean*. The first part rows list the performances for the three baselines, the middle three parts display active learning query strategies used previously for NER, and the last two parts present the results of the proposed active learning strategies. In this table, for each dataset results from the last four iterations before convergence are shown and performances are compared at the same cost, where cost is measured by the number of sentences annotated.

AL Methods	CoNLL-03				BC5CDR				NCBI-Dz.			
	iteration 8	iteration 9	iteration 10	iteration 11	iteration 6	iteration 7	iteration 8	iteration 9	iteration 6	iteration 7	iteration 8	iteration 9
	%4	%9	%17.5	%35	%1.5	%3	%6	%11	%5	%10	%19	%39
RS	76.5±0.2	79.0±0.8	81.7±2.5	84.7±3.4	66.9±0.5	69.9±0.5	73.7±0.3	79.1±0.1	61.1±1.1	64.0±0.9	69.5±0.5	78.2±0.2
LSS	66.3±0.4	70.1±0.5	73.6±0.4	74.3±0.6	62.6±0.4	65.2±0.7	73.4±0.3	79.8±0.1	51.8±0.7	61.6±0.6	71.3±0.5	78.6±0.3
PAS	65.9±0.9	69.5±0.8	82.7±0.3	87.2±0.2	54.6±1.8	59.4±1.0	66.2±0.8	75.1±0.8	57.5±0.7	61.7±0.8	69.7±0.9	78.2±0.4
sAP	78.1±0.4	85.7±0.2	90.1±0.2	92.5±0.1	71.0±0.3	72.2±0.5	78.5±0.1	82.6±0.1	61.2±1.0	68.0±0.8	76.3±0.4	81.8±0.3
sTE	76.9±0.4	84.3±0.3	89.9±0.1	92.6±0.1	70.5±0.5	73.1±0.5	78.1±0.3	82.6±0.1	61.5±1.0	68.7±1.1	76.7±0.5	81.6±0.2
sTP	76.2±0.4	84.4±0.2	89.9±0.1	92.5±0.1	70.8±0.3	73.3±0.3	78.4±0.2	82.6±0.2	59.5±1.9	69.0±1.4	76.9±0.7	81.1±0.3
sTM	78.8±0.3	85.2±0.3	90.3±0.0	92.7±0.1	71.0±0.4	72.5±0.3	77.9±0.3	82.5±0.1	60.3±1.4	67.9±1.1	76.6±0.6	81.5±0.2
$\mu(\texttt{single})$[a]	77.5±0.4	84.9±0.3	90.1±0.1	92.6±0.1	70.8±0.4	72.8±0.4	78.2±0.2	82.6±0.1	60.6±1.4	68.4±1.1	76.6±0.6	81.5±0.3
tAP	78.6±0.4	86.4±0.2	90.8±0.1	92.8±0.0	70.1±0.4	75.2±0.5	80.2±0.2	83.5±0.1	59.4±1.1	71.0±0.7	77.1±0.7	81.5±0.4
tTE	78.2±0.3	86.3±0.3	90.8±0.1	92.8±0.1	68.8±0.5	75.4±0.5	80.4±0.1	83.5±0.1	59.1±1.1	70.1±0.8	76.8±0.8	81.0±0.4
tTP	78.2±0.6	86.5±0.1	90.8±0.2	92.7±0.1	68.6±0.6	75.0±0.2	80.2±0.1	83.5±0.1	61.0±1.1	71.7±0.5	76.2±0.4	81.6±0.3
tTM	78.1±0.6	87.0±0.1	90.6±0.2	92.8±0.1	68.6±0.6	75.4±0.2	80.4±0.1	83.4±0.1	61.4±1.1	70.7±0.5	76.5±0.4	81.3±0.3
$\mu(\texttt{total})$[b]	78.3±0.4	86.5±0.2	90.8±0.1	92.8±0.1	69.0±0.6	75.2±0.4	80.3±0.1	83.5±0.1	60.2±1.1	70.9±0.7	76.7±0.6	81.3±0.4
nAP	65.0±0.9	75.0±0.6	87.0±0.2	92.1±0.1	69.5±0.6	71.3±0.5	76.2±0.4	81.9±0.1	55.5±1.2	60.8±1.1	72.5±0.7	79.9±0.3
nTE	61.1±1.3	74.7±0.6	81.2±5.2	91.7±0.2	68.9±0.7	71.9±0.4	77.4±0.2	82.1±0.2	55.1±2.0	58.7±1.2	72.1±0.6	79.8±0.3
nTP	64.3±1.6	75.6±0.6	86.8±0.2	92.0±0.1	68.7±0.4	71.6±0.3	77.0±0.2	81.8±0.1	55.1±3.0	61.3±1.0	72.4±0.7	80.6±0.4
nTM	64.7±1.0	74.7±0.8	86.6±0.2	84.6±5.1	69.9±0.5	71.4±0.5	76.6±0.2	82.0±0.1	53.4±1.6	59.1±1.7	73.0±0.5	79.4±0.5
$\mu(\texttt{normalized})$[c]	63.8±1.2	75.0±0.7	85.4±2.6	90.1±2.6	69.2±0.6	71.6±0.4	76.8±0.3	82.0±0.1	54.8±2.1	60.0±1.3	72.5±0.6	79.9±0.4
tpAP	79.8±0.4	86.6±0.2	90.6±0.1	92.7±0.1	69.6±0.7	74.8±0.3	79.9±0.3	83.0±0.1	60.6±1.0	70.2±1.0	77.6±0.5	81.1±0.3
tpTE	79.2±0.3	86.6±0.3	90.7±0.1	92.6±0.1	69.2±0.4	74.8±0.4	79.8±0.2	83.2±0.1	59.1±1.0	69.9±1.1	76.9±0.6	81.8±0.2
tpTP	79.2±0.5	86.6±0.2	90.7±0.2	92.6±0.1	69.7±0.4	74.4±0.4	79.7±0.2	83.0±0.2	61.8±1.0	70.8±0.8	76.7±0.5	81.8±0.3
tpTM	78.5±0.5	86.0±0.2	90.1±0.2	92.3±0.1	70.4±0.4	73.9±0.4	78.6±0.2	82.6±0.2	61.2±1.0	68.3±0.8	76.2±0.5	80.5±0.3
$\mu(\texttt{total-pos})$[d]	79.2±0.4	86.4±0.2	90.5±0.2	92.6±0.1	69.7±0.5	74.4±0.4	79.5±0.2	83.0±0.2	60.7±1.2	69.8±1.0	76.8±0.6	81.3±0.3
dpAP	76.0±0.5	84.2±0.1	89.7±0.1	92.4±0.1	70.6±0.5	72.5±0.6	77.0±0.2	82.0±0.1	59.3±2.1	65.2±0.7	76±0.6	81.3±0.2
dpTE	74.7±0.6	84.2±0.4	90.0±0.1	92.4±0.1	70.6±0.5	72.6±0.7	77.3±0.3	82.1±0.1	57.2±1.7	65.2±1.3	75.1±0.7	81.3±0.4
dpTP	76.5±0.3	84.7±0.2	89.7±0.1	92.5±0.1	71.5±0.5	73.3±0.3	77.2±0.3	81.9±0.2	53.6±2.0	67.9±0.7	75.2±0.4	80.8±0.4
dpTM	75.7±0.4	84.1±0.2	89.6±0.1	92.1±0.1	70.7±0.4	73.4±0.4	77.1±0.3	81.8±0.1	58.5±1.8	66.4±1.3	74.8±0.5	80.9±0.6
$\mu(\texttt{dnorm-pos})$[e]	75.7±0.5	84.3±0.2	89.8±0.1	92.4±0.1	70.8±0.5	73.0±0.5	77.2±0.3	82.0±0.1	57.2±1.9	66.2±1.0	75.3±0.6	81.1±0.4

[a] Average F_1-score of sTE, sTP, sTM and sAP. [b] Average F_1-score of tTE, tTP, tTM and tAP. [c] Average F_1-score of nTE, nTP, nTM and nAP.
[d] Average F_1-score of tpTE, tpTP, tpTM and tpAP. [e] Average F_1-score of dpTE, dpTP, dpTM and dpAP.



(a) CoNLL-03

(b) BC5CDR

(c) NCBI-Disease

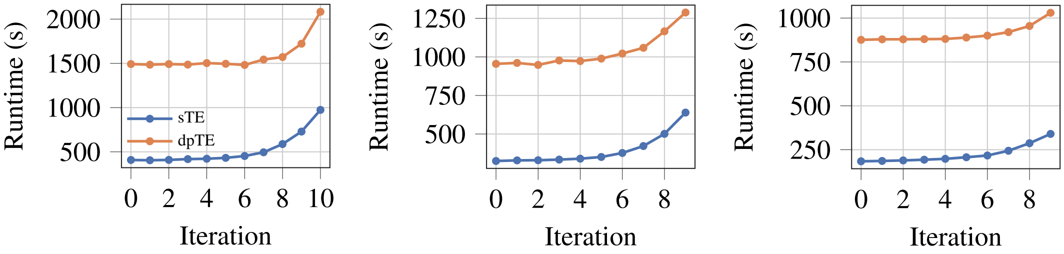
Figure S1. Average F_1-scores of different aggregation methods and the baselines with respect to the total number of annotated tokens.

Appendix S2.

Table S2. F_1-score values achieved in the last four iterations before convergence by the tpTE method on CoNLL-03 corpus. Three different outlier detection alternatives are compared: Global-Local Outlier Score from Hierarchies, Local Outlier Factor, and none (i.e., no outlier detection).

	CoNLL-03			
Outlier Method	**iteration 8**	**iteration 9**	**iteration 10**	**iteration 11**
	%4	%9	%17.5	%35
GLOSH[a]	72.0	80.6	86.4	90.7
LOF[b]	66.5	75.1	81.3	86.6
None[c]	66.0	73.5	83.3	87.3

[a] Global-Local Outlier Score from Hierarchies
[b] Local Outlier Factor [c] No Outlier Detection



(a) CoNLL-03 **(b)** BC5CDR **(c)** NCBI-Disease

Figure S2. A representative runtime comparison between active learning iterations of sTE and dpTE. Total runtime includes both the training time and the query.