12.12-MPQA语料统计结果

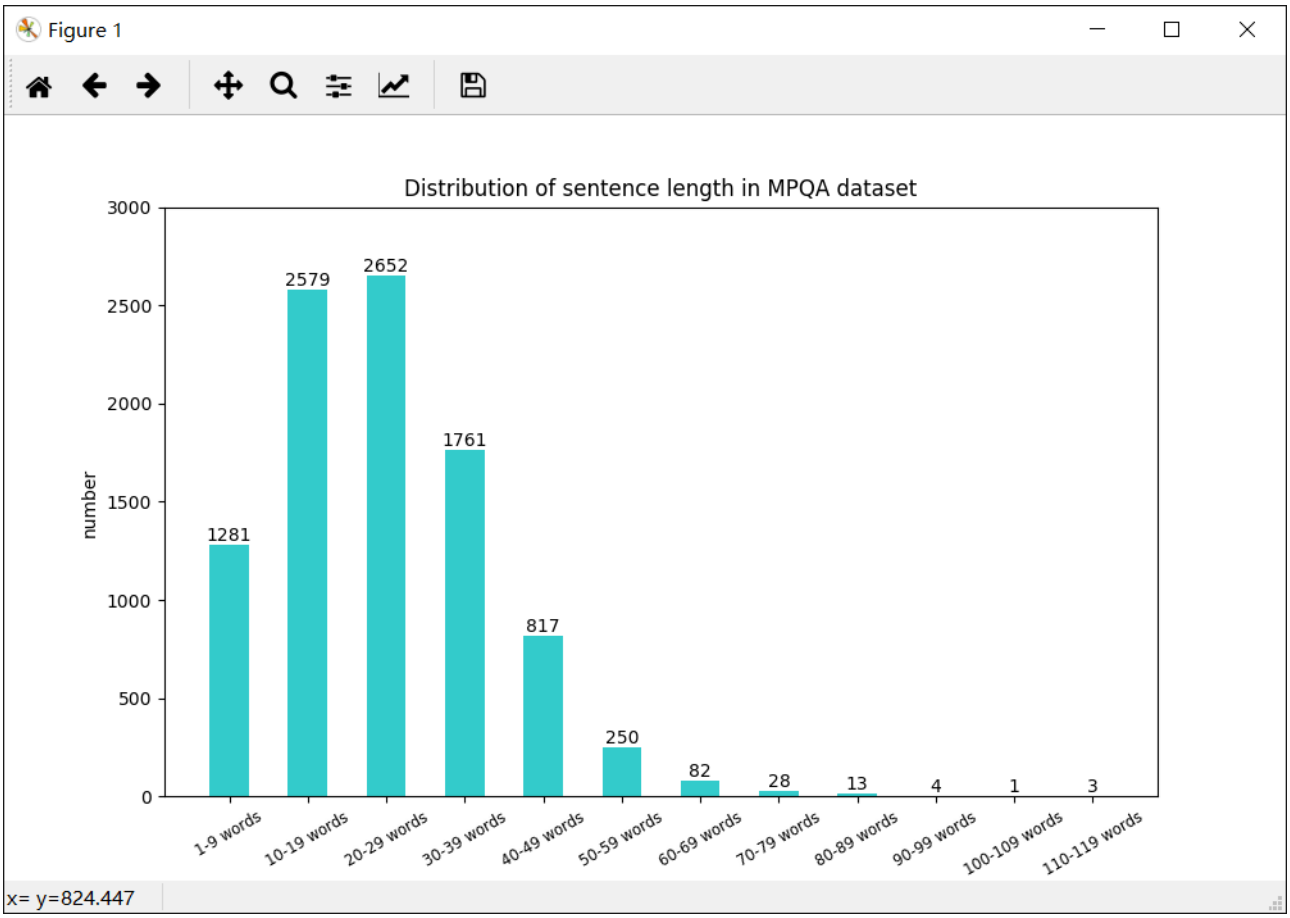
MPQA语料统计

1. 统计语料中句子的长度分布，分布情况如下图所示：

表 1 语料中句子长度分布表

length	1~9	10~19	20~29	30~39	40~49	50~59	60~69	70~79	80~89	90~99	100~109	110~119	Total
number	1281	2579	2652	1761	817	250	82	28	13	4	1	3	9471

图 1 语料中句子长度分布



2. 统计语料中各实体的长度分布，各实体分布情况和实体之间的对比情况如下图所示：

表 2 AGENT实体的长度分布表

length	1~2	3~5	6~8	9~11	12~14	15~17	18~20	21~23	24~26	27~29	30~32	33~35	36~38	Total
number	3153	858	175	89	28	14	4	3	1	1	0	1	2	4329

图 2 AGENT实体的长度分布

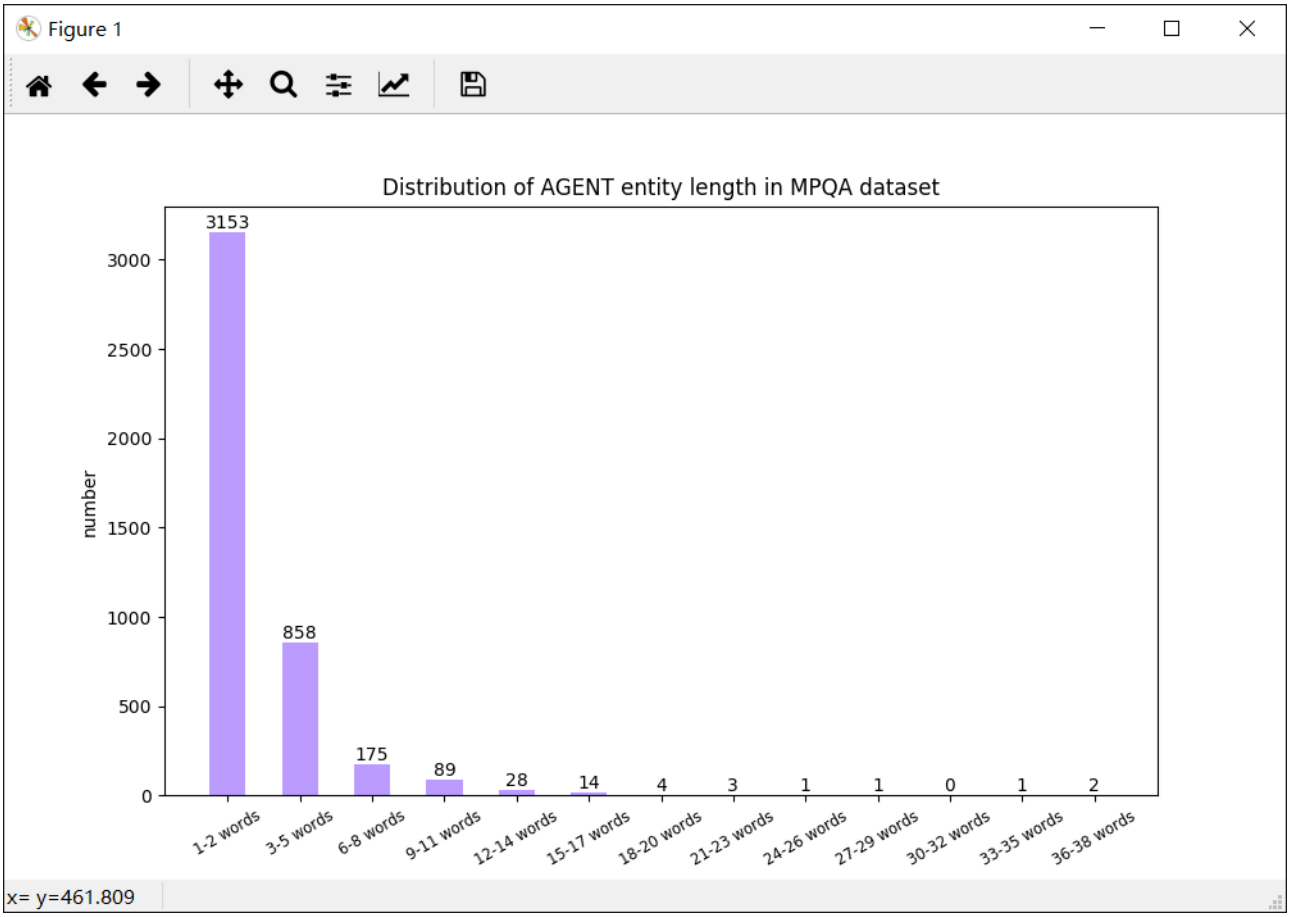


表 3 TARGET实体的长度分布表

length	1~4	5~9	10~14	15~19	20~24	25~29	30~34	35~39	40~44	45~49	50~54	55~59	Total
number	2856	1167	437	201	84	32	13	3	3	4	0	1	4801

图 3 TARGET实体的长度分布

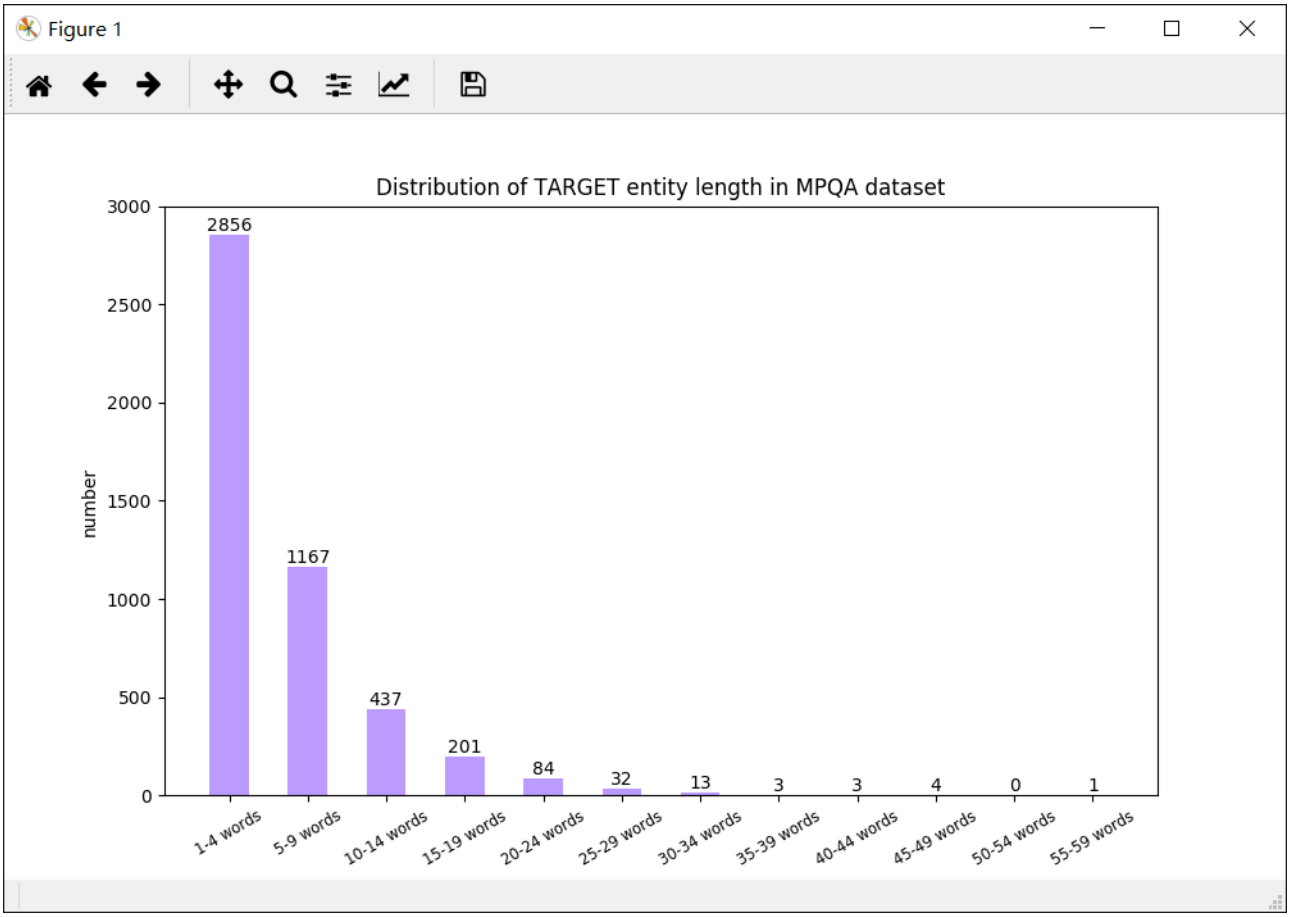


表 4 DSE实体的长度分布表

length	1	2	3	4	5	6	7	8	9	10	11	14	15	18	34	Total
number	3452	1260	705	287	128	52	34	15	6	1	4	2	3	1	1	5951

图 4 DSE实体的长度分布

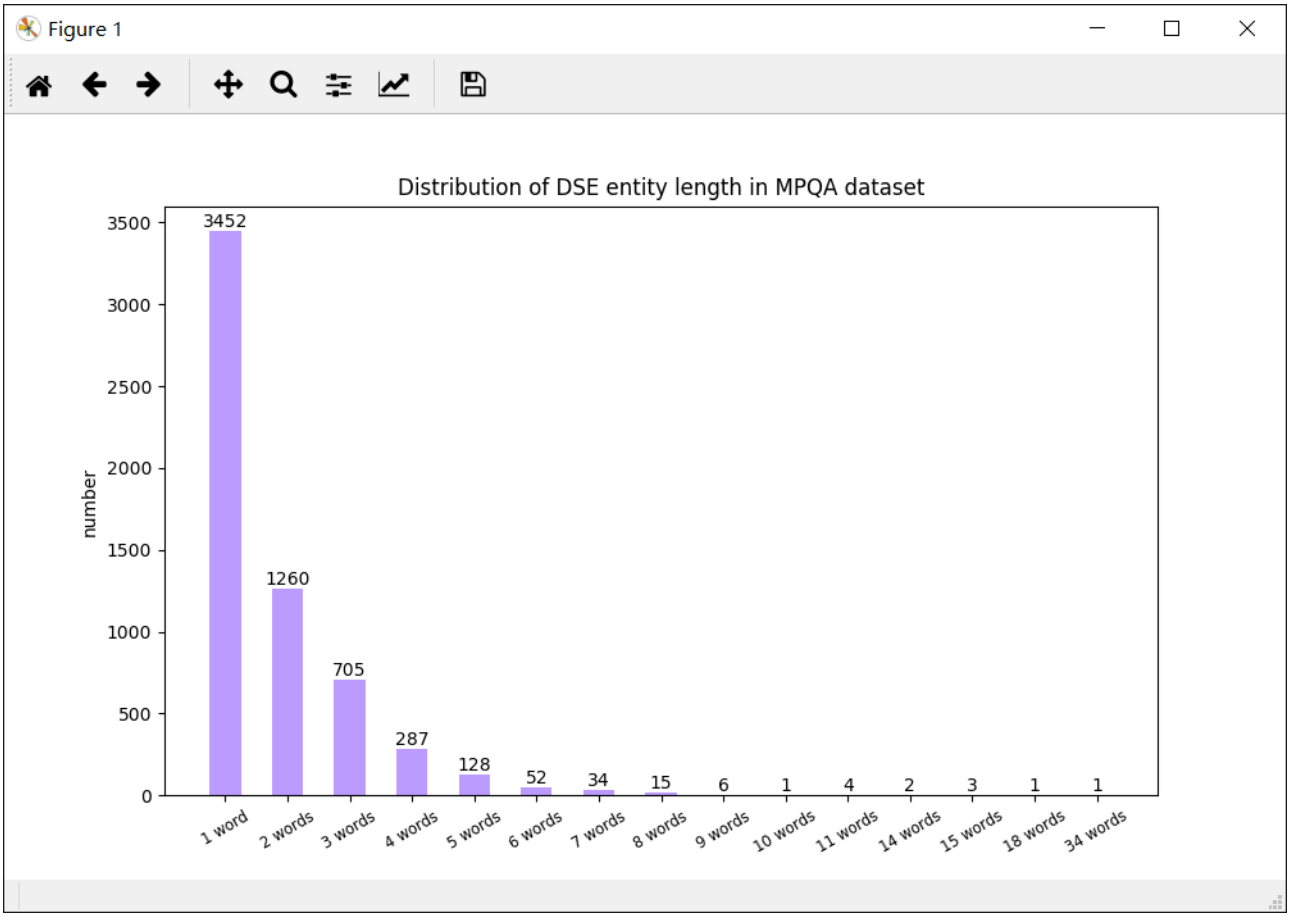
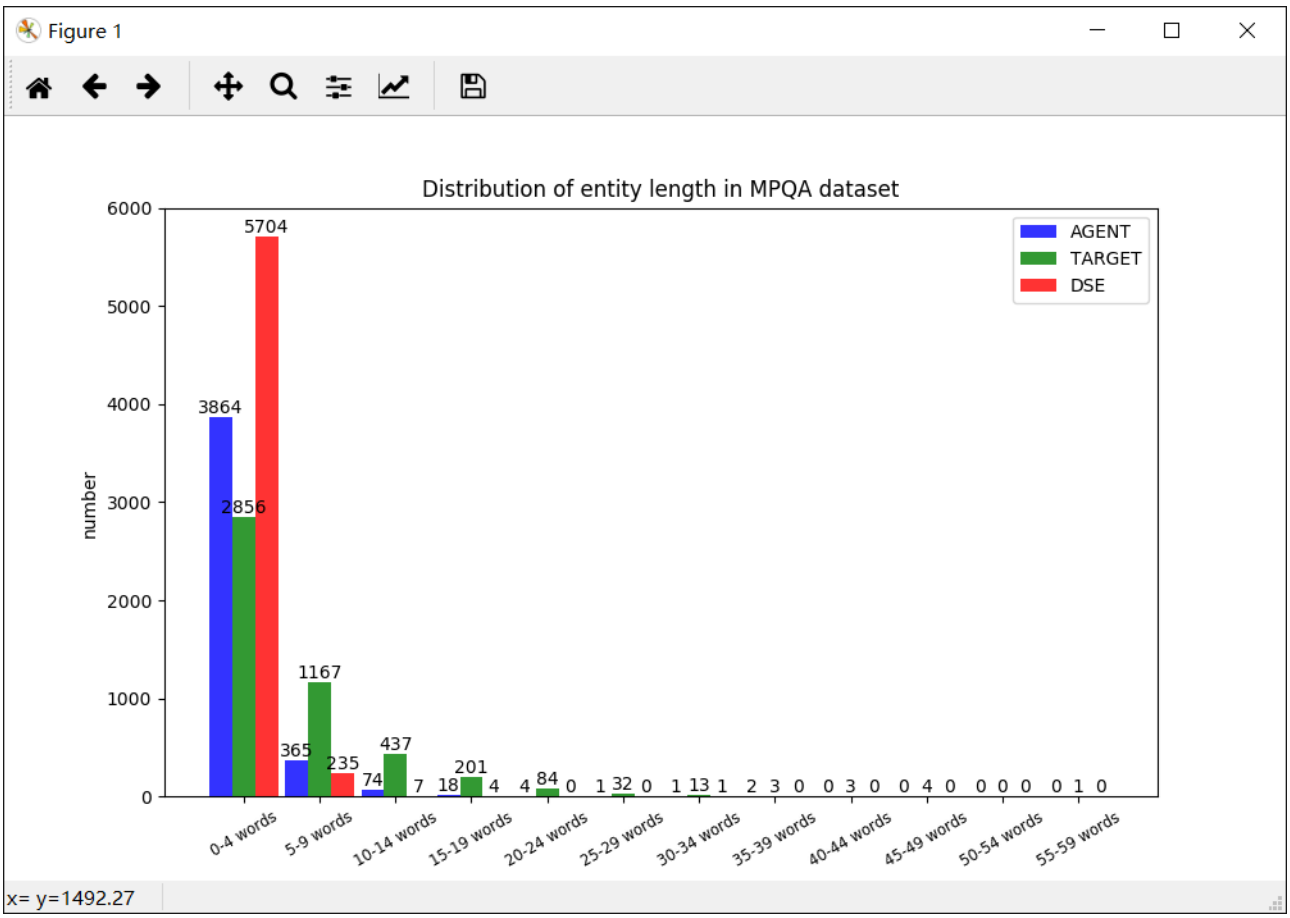


图 5 各实体的长度分布对比图



（1）验证：

以上数据与之前统计的实体个数是相吻合的，统计结果应该是正确的。

表 5 all_ILP实体统计表

	Lines	Sentences	Total	AGENT	DSE	TARGET
bio	236970	9471	15081	4329	5951	4801
bmes	236970	9471	15081	4329	5951	4801
old_version		9471	15080	4329	5951	4800

3. 统计语料中AGENT，TARGET，DSE之间数量上的对应关系

表 6 统计每个DSE对应的AGENT和TARGET实体个数

DSE	0	1	2	3	4	Total
AGENT	1169	4752	6	0	0	5927
TARGET	1252	4424	231	18	2	5927

表 7 统计每个AGENT对应的DSE实体个数

AGENT	1	2	3	4	Total
DSE	3926	373	28	2	4764

表 8 统计每个TARGET对应的DSE实体个数

TARGET	1	2	3	Total
DSE	4655	145	1	4948

（1）验证：

表 6 中4752+2*6+4424+231*2+18*3+2*8=9712

表7与表8之和：4764+4948=9712

这与之前统计的数据是一致的，统计结果应该是正确的。

表 9 all_ILP抽取关系后行数变化统计表

	all_ILP lines	lines-add rel	add lines	add pairs
number	236970	246682	9712	9712

（2）将统计结果绘制成直方图，如下所示：