**Similar pair 实验报告**

1. 目的

现有的同花顺金融领域语言模型在训练时训练集缺少一些金融领域英文缩写，所以这个模型对测试集中的一些英文缩写识别效果较差。现在通过similar pair方法，在原语言模型中加入英文缩写，生成新的语言模型，并对比测试结果，评估similar pair在此问题上的有效性。

1. 过程

从测试集中找出所有的金融领域英文缩写，再在baseline语言模型中找到合适的英文缩写作为position，然后指定weight值，生成加入了英文缩写的语言模型。调节weight值并测试，找出最优的weight值，从而得到最终语言模型。

1. 英文缩写列表

英文缩写列表是测试集中所有金融领域英文缩写的列表，将被转换成语言模型加入到baseline语言模型中，从而提高这些词被识别的概率。

测试集是/work6/weiy/work/lm\_training/test\_finance\_with\_en，这里的金融领域英文缩写主要是股票投资指标。首先提取测试集中所有英文缩写（大写），然后参考上下文，剔除明显不是股票投资指标的英文缩写（如“CEO杨元庆”），得到英文缩写列表。

1. 找position

Position是在baseline语言模型中人为选定的词，选好后，英文缩写列表的语言模型将被插入到baseline语言模型中position所在的所有位置，并借用position的相应weight值。根据经验，position通常选择一个词。

因此，position词应该在baseline中经常出现，本身没有歧义（如“AR指标”和“AR游戏”），左右邻近词丰富，能覆盖测试集中股票投资指标的多数上下文情境。根据以上原则，position被确定为“KDJ”。

1. 造发音词典

测试新语言模型时，由于引入新词，所以需要修改发音词典。首先参照老词表列出英文缩写列表中的新词，将它们注音，然后与similar pair工具生成的lex.txt合并，得到新发音词典。

1. 确定weight

将英文缩写列表语言模型插入到baseline语言模型中时需要指定weight值，英文缩写语言模型在baseline语言模型某处的实际weight值由这个weight值和此处position的weight值共同决定。Weight值越小，英文缩写被识别的概率越高，但其他词被误识别的概率也会增高。所以，可以通过实验，调节weight值，找到一个最优值。

1. 结果

|  |  |
| --- | --- |
| lm | wer |
| fin\_all\_nv\_3gram\_1e-7 | 31.66 |
| fin\_all\_nv\_abbtest\_3gram\_1e-7 | 29.67 |

Table 1 Similar pair LM and baseline

|  |  |
| --- | --- |
| weight | wer |
| 0 | 29.67 |
| -100000000 | 159.03 |
| -30 | 398.9 |
| -20 | 100.96 |
| -12 | 28.79 |
| -11 | 27.04 |
| -10 | 25.97 |
| -9 | 25.62 |
| -8 | 25.36 |
| -7 | 25.55 |
| -5 | 26.80 |
| -1 | 29.17 |
| 100000000 | 31.66 |

Table 2 Didderent weights

|  |  |  |
| --- | --- | --- |
| 1 | baseline | NB 给 |
| 2 | similar pair | NBA |
| 3 | 测试集 | NBA |
| 4 | baseline | 认为 ECD |
| 5 | similar pair | MACD |
| 6 | 测试集 | MACD |
| 7 | baseline | 二 B 安 |
| 8 | similar pair | AI 比 一 |
| 9 | 测试集 | IBM |
| 10 | baseline | ORB 二 指标 的 使用 集中 在 阿 比亚 曲线 所 处 位置 和 运行 方向 两 方面 |
| 11 | similar pair | ARBR 指标 的 使用 集中 在 ARBR 曲线 所 处 位置 和 运行 方向 两 方面 |
| 12 | 测试集 | ARBR指标的使用集中在ARBR曲线所处位置和运行方向两方面 |
| 13 | baseline | DBO 由 惠特曼 八 而 沙特 提出 了 |
| 14 | similar pair | DPO 有 惠特曼 八 而 沙特 提出 了 |
| 15 | 测试集 | DPO由惠特曼巴塞特提出了 |
| 16 | baseline | 一 安危 指标 是 有 一 条 一 百 五 一线 和 一 条 均线 组成 |
| 17 | similar pair | 一 按 五 一 指标 是 有 一 条 一 百 五 一线 和 一 条 均线 组成 |
| 18 | 测试集 | EMV指标是由一条EMV线和一条均线组成 |
| 19 | baseline | 如果 欧 系 元 零零 线 其 价格 远离 平 宁津县 应 注意 价格 很 可能 向 平均线 的 回归 |
| 20 | similar pair | 如果 OSC 元 零零 线 其 价格 远离 平 EUR 均 现 应 注意 价格 很 可能 向 平均线 的 回归 |
| 21 | 测试集 | 如果OSC远离零线即价格远离平均线应注意价格很可能向平均线回归 |
| 22 | baseline | 城 口 历史 走势 确定性 小 达到 席位 一 指标 及 小吃 的 标准 |
| 23 | similar pair | 城 口 历史 走势 确定性 小 达到 CYE 指标 及 小吃 的 标准 |
| 24 | 测试集 | 参考历史走势确定一下达到CYE指标极小值的标准 |
| 25 | baseline | 俺们 点 二 十 迈 科 支撑 压力 指标 的 中期 压力线 |
| 26 | similar pair | AM 点 二 十 迈 科 支撑 压力 指标 的 中期 压力线 |
| 27 | 测试集 | MIDR是麦克支撑压力指标的中级压力线 |
| 28 | baseline | 当 股价 处于 盘整 阶段 四 依据 大约 一 厅 二 与 暗暗 第 二 构成 拿出 区域 |
| 29 | similar pair | 当 股价 处于 盘整 阶段 四 依据 WAD 给 二 与 MIDR 构成 拿出 区域 |
| 30 | 测试集 | 当股价处于盘整阶段时依据WEKR与MIDR构成卖出区域 |
| 31 | baseline | 跌 没 指标 也 包括 日 跌 美 指 表 周 跌 美 指标 等 各种 类型 |
| 32 | similar pair | DMI 指标 也 包括 日 跌 美 指 表 周 跌 美 指标 等 各种 类型 |
| 33 | 测试集 | DMA指标也包括日DMA指标周DMA指标等各种类型 |
| 34 | baseline | 五 压制 的 丰富 不同 给出 的 信号 也 不同 |
| 35 | similar pair | VR 之 的 丰富 不同 给出 的 信号 也 不同 |
| 36 | 测试集 | VR值的分布不同给出的信号也不同 |
| 37 | baseline | B 二 百 一 十 六 一 二 是 指 六日 平均 数值 与 十二日 平均 数值 之间 的 差 |
| 38 | similar pair | BIAS 六 一 二 是 指 六日 平均 数值 与 十二日 平均 数值 之间 的 差 |
| 39 | 测试集 | BIAS六一二是指六日平均数值与十二日平均数值之间的差 |
| 40 | baseline | DBS 对比 强弱 指标 十 个 短线 指标 |
| 41 | similar pair | DBQR 对比 强弱 指标 十 个 短线 指标 |
| 42 | 测试集 | DBQR就是对比强弱指标是一个短线指标 |
| 43 | baseline | 新闻 ICE 指标 就是 市场 盈亏 指标 |
| 44 | similar pair | CYS 指标 就是 市场 盈亏 指标 |
| 45 | 测试集 | CYS指标就是市场盈亏指标 |
| 46 | baseline | 股票 ST O 开始 指标 注释 及 应用 |
| 47 | similar pair | 股票 开始 T OS 指标 注释 及 应用 |
| 48 | 测试集 | 股票STOS指标注释及应用 |
| 49 | baseline | 低点 每 平 均 超 指标 处于 趋向 类 指标 也 是 一 种 趋势 分析 指标 |
| 50 | similar pair | D 每 平 均 超 指标 处于 趋向 类 指标 也 是 一 种 趋势 分析 指标 |
| 51 | 测试集 | DMA平均差指标属于趋向类指标也是一种趋势分析指标 |
| 52 | baseline | 谁 最 低 延时 走上 了 而 股价 K线 连续 几 天 还是 走 平 的 上下 波动 幅度 非常 小 |
| 53 | similar pair | 谁 最 低 而是 走上 了 而 股价 K线 连续 几 天 还是 走 平 的 上下 波动 幅度 非常 小 |
| 54 | 测试集 | CJDX走上去了而股价K线连续几天还是走平的上下波动幅度非常小 |
| 55 | baseline | 输出 阿曼 西奥 求职 为 求 |
| 56 | similar pair | 输出 AM 为 求 |
| 57 | 测试集 | 输出MACHO值为CHO |
| 58 | baseline | 三 是 股价 创新 低 而 第一 批 区 职位 创新 低 数 底 背离 |
| 59 | similar pair | 三 是 股价 创新 低 而 DPO 职位 创新 低 数 D 背离 |
| 60 | 测试集 | 三是股价创新低而DPO值未创新低属底背离 |

Table 3 Results contrast

Note: 1. fin\_all\_nv\_3gram\_1e-7是baseline语言模型，由59G金融领域数据训练而成。

2. fin\_all\_nv\_abbtest\_3gram\_1e-7是weight值为0的similar pair语言模型。

3. 英文缩写列表是/work6/weiy/work/lm\_training/abbreviation\_test\_finance。

4. AM为/work5/release/project/myhexin/am/v0.1。

5. 测试集是/work6/weiy/work/lm\_training/test\_finance\_with\_en。

6. fin\_all\_nv\_3gram\_1e-7发音词典为/work6/weiy/work/lm\_training/lexicon.v0417。

7. fin\_all\_nv\_abbtest\_3gram\_1e-7发音词典为/work6/weiy/work/similar\_pair/test/G\_merge/ext\_THS\_test/lex\_all.txt。

8. position为”KDJ”。

1. 结论

使用similar pair扩充语言模型可以在指定测试集上提升识别率。在金融领域英文缩写的识别中，通过调节similar pair的weight值，当weight值取-8时，识别率最高。性能提升幅度较大（wer从31.66%降低到25.36%）。因此，可以将similar pair作为在语言模型中加入新词，提高识别率的手段。